YAŞAR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

MASTER THESIS

AUTOMATED DATA FEEDBACK FOR REAL TIME SCHEDULING

FIRAT YILMAZ

THESIS ADVISOR: ASST.PROF. ADALET ONER

INDUSTRIAL ENGINEERING

PRESENTATION DATE: 25.11.2019

BORNOVA / İZMİR
NOVEMBER 2019

We certify that, as the jury, we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Jury Members:	**Signature:**
Asst. Prof. Adalet ÖNER Yaşar University	…………………
Asst. Prof. Erdinç ÖNER Yaşar University	…………………
Asst. Prof. Erkan KABAK Yaşar University	…………………

Prof. Cüneyt GÜZELİŞ
Director of the Graduate School

ABSTRACT

AUTOMATED DATA FEEDBACK FOR REAL TIME SCHEDULING

Fırat Yılmaz

MSc, Industrial Engineering

Advisor: Asst.Prof.Dr. Adalet Öner

Co-Advisor: Assoc.Prof.Dr. Gökhan Kılıç

November 2019

This study involves in automated data feedback for real time job shop scheduling algorithms in the industrial environment. It is an application of horizontal integration concept that is one of the building blocks of the Industry 4.0 philosophy. The thesis considers an application in a company that produces printed packaging films and printed paper bags. The manufacturing process follows a classical job-shop production environment. While scheduling the orders, the management tries to find an optimal solution, which is a balance between two performance measures. The first one is customer satisfaction that is represented by due dates, whereas the second one is the setup times that incur opportunity costs. For scheduling purpose, Giffler & Thompson scheduling algorithm has been considered since it is widely known and easy to implement. In usual scheduling process, the values of the input parameters are given into the algorithm and a solution is obtained. However, in practice, some changes may occur in the values of input parameters in planning horizon due to some reasons such as machine failures, unpredictable shortages in workforce, fluctuations in the efficiencies of the workers due to employee turnovers etc. In this study, it is intended to show how to update or modify production schedules dynamically using real-time data collected from the machinery.

Key Words: Job Shop Scheduling, Manufacturing Execution Systems, Enterprise Resource Planning, Big Data

ÖZ

OTOMATIK VERI GERI BILDIRIM ILE ANLIK ÇIZELGELEME

Fırat Yılmaz
Yüksek Lisans
Danışman: Dr.Öğr.Üyesi Adalet Öner
Yardımcı Danışman: Doç.Dr.Gökhan Kılıç
Kasım 2019

Bu tezde, endüstriyel ortamda, atölye tipi çizelgeleme algoritmalarının kullanılması için gerçek zamanlı otomatik veri besleme konusu çalışılmıştır. Bu konu aslında Endüstri 4.0 felsefesinin yapı taşlarından biri olan yatay entegrasyon çalışmalarına bir örnek teşkil etmektedir. Bu çalışmada baskılı ince paketleme filmleri ve baskılı kâğıt poşetler üreten bir işletmedeki uygulamalar ele alınmıştır. İşletmede imalat süreci atölye tipi üretim ortamında yürütülmektedir. Üretim çizelgelemesi yapılırken siparişlerin teslim zamanlarına uyum (müşteri memnuniyeti) ve makine hazırlık süreleri (maliyet) arasında bir denge kurulmaya çalışılmaktadır. Çizelgeleme için Giffler ve Thompson algoritması dikkate alınmıştır. Bu algoritmada önceden belirlenmiş girdi parametreleri kullanılarak bir çözüm ortaya konulur. Ancak uygulamada arızalar, öngörülemeyen personel eksikleri, personel devir hızından dolayı işçi performansındaki dalgalanmalar veya planlama bölümüne yansımayan yerel mikro iyileştirmeler gibi nedenlerle girdi parametrelerinin değerlerinde değişimler meydana gelebilmektedir. Bu çalışmada, üretimde kullanılan makine ve donanımlardan gerçek zamanlı anlık veriler toplanabildiği şartlarda girdi parametrelerinin güncellenerek yeni çizelgelerin dinamik bir şekilde nasıl yenilendiği gösterilmektedir.

Anahtar Kelimeler: Atölye Tipi Üretim Çizelgeleme, Üretim Yönetim Sistemleri, Kurumsal Kaynak Planlama, Büyük Veri

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor Adalet Öner and co-advisor Gökhan Kılıç for their guidance and patience during this study.

I am also grateful to my dear parents who encouraged me to start my master degree programme and the motivations that gave me during all the process, also many regards to my dear wife for his moral support in all processes and understanding me to be able to complete this difficult and important step for my life. I would like to thank to Baran Ambalaj to support me during all process of my thesis.

Fırat Yılmaz

İzmir, 2019

TEXT OF OATH

I declare and honestly confirm that my study, titled "Automated Data Feedback for Real Time Scheduling" and presented as a Master's Thesis, has been written without applying to any assistance inconsistent with scientific ethics and traditions. I declare, to the best of my knowledge and belief, that all content and ideas drawn directly or indirectly from external sources are indicated in the text and listed in the list of references.

Fırat YILMAZ

....

November 25, 2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

.SYMBOLS AND ABBREVIATIONS

ABBREVIATIONS:

SPT Shortest Processing Time

LST Least Set Up Time

PIF Product Information Form

EDD Earliest Due Date

LSNEDD Least Set Up Time and Earliest Due Date

MES Manufacture Execution System

ERP Enterprise Resource Planning

BOM Bill of Material

WIP Work in Process

ROU Route of Material

JSP Job Shop Scheduling Problems

FJSP Flexible Job Shop Scheduling Problems

CHAPTER 1
INTRODUCTION

There are many variables in packaging industry. Although the processes are very sensitive, the type and quality of the material used have a great importance. Obviously, each product has its own unique design. Even customers bring their special designs so the packaging company has many types of designs and there are many parameters in its system. In a sector with such variability, the fact that most of the raw materials are imported from abroad adds one more level to degree of variability in the production environment. In addition to such variability, the product range in the printing industry is vast. Each BOM has very different routes. This means that many different production lines are employed for each product. If the production schedules are not developed efficiently, it easily leads to some problems such as high waiting times of semi-finished products on production lines, increase of WIP stocks and frequent failures in meeting the customers' deadlines.

In this complexity of the packaging industry, production schedules are developed manually in the sponsor company. Each job is assigned to the appropriate machines by the production planning engineers after the resource requirements have been calculated and estimated how long it will wait in the next process. Especially the averagely 200 minutes set up times of the packaging industry increase the inefficiency here. While the average set up time in the printing process is 180 minutes, this time is 600 minutes in bag production and every wrong decision to be made at this point can cause an increase in set up times. In this respect, there is a risk in complying two of the basic expectations of customer, which are cost and delivery time (due date). Setup times will increase the costs and cause deviation from due date. At this point, a systematic and robust scheduling process should be employed in order to develop efficient production schedules. Thanks to developing technology, companies now manage their processes with the help of ERP. ERP acts as the nervous system of companies in many processes such as accounting, human resources, production and logistics processes.

Within the company, the aim of the company to use a scheduling algorithm to obtain much more efficient production schedules than the schedules prepared manually. Furthermore, it is desirable to run this algorithm in collaboration with the current ERP software used in the company.

In the company, the manufacturing process follows a classical job-shop production environment. Although there are some mathematical models for job-shop scheduling problems, it is not practical to use them to solve real life problems since it takes too much time (it is known that the job shop scheduling problem is NP-hard). For this reason, many heuristic methods have been proposed to obtain near optimal solutions in a way shorter times. One of those algorithms is the Giffler and Thompson (G&T) algorithm [2].

On the other hand, the production process is dynamic. Although the data collected here is processed into the ERP environment, it remains cumbersome for daily, weekly or monthly actions. At this point, "Manufacturing Execution Systems" (MES), which entered into our lives with industry 4.0, shows us the dynamism of the production process instantly. MES is an information system to monitor production environment and give us real time data for realize the inefficiency in production lines. MES connects all of your machines in one environment also give an ability to monitor and control them. MES collect data from machine, robots or workers with help of sensors. This data allows us to understand factory settings and optimize them correctly. In the light of this information, we can monitor our production environment instantly and make dynamic decisions.

ERP systems are not good software for scheduling purposes since their scheduling modules have generic designs and can't be easily used in particular industries. However, external algorithms such as Giffler and Thompson algorithm can be integrated into the ERP software. Even in that case, it is not possible (at least in current configuration in the company) to use real time data collected from the sensors. Therefore, they have problems keeping up with the ever-changing structure of the production line and hence the schedules generated by ERP scheduling module are useless since the values of input parameters are changing dynamically.

For this reason, it is desired to establish a communication link between the real time database (raw MES data collected from the production environment) and the ERP database that holds the input parameters of the Giffler and Thompson (G&T) algorithm. Data from MES database are to be summarized with various data mining methods and then to be fed into ERP system. The ideal structure should be as shown in Figure 1.1. In order to establish the ideal structure, the ERP system has to be up and running in an efficient way. Then MES Manufacture Execution System) should be installed and integrated into ERP system which is called the "horizontal integration".

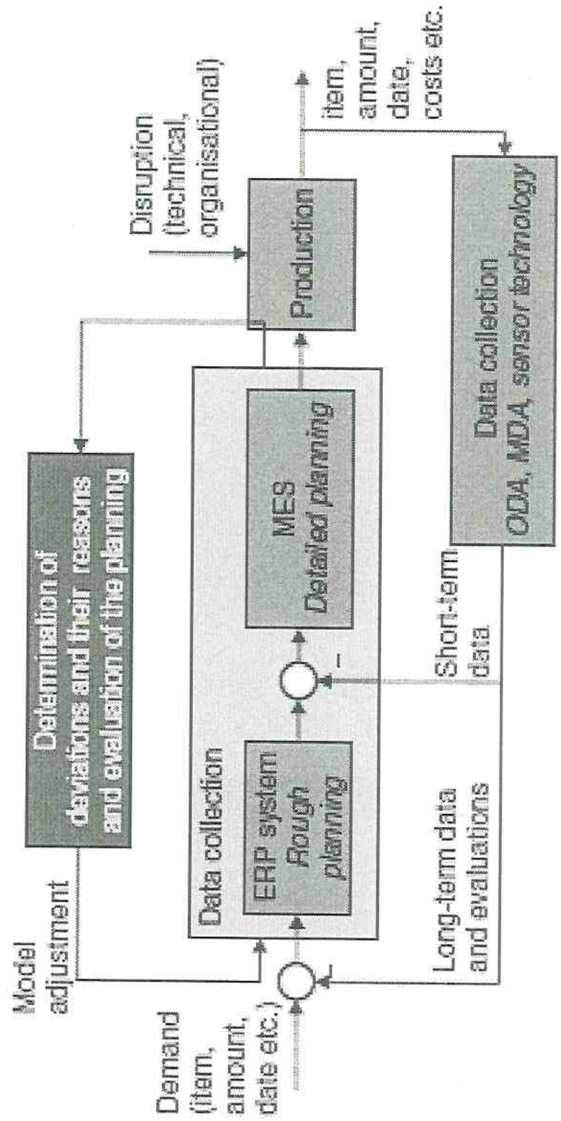


Figure 1.1. MES and ERP Integration for Real Time Scheduling

The sponsor company has a facility that produces printed bags for packaging various commodities. There are several machines used in production process. Each machine is devoted to a special operation and each product has its own route through the machines. The production environment resembles the job shop problem. A bag production within the company can enter one of the printing, lamination, slitting operations or directly to the assembly operation without any prior operation. In the company, the planning analysts prepare the production schedules manually using the offline spreadsheet application (MS Excel). The efficiency of the schedule is in question. Additively, there is always a probability of making mistakes due to manual data management and calculations. Furthermore, the values of input parameters of the scheduling problem changes over time and those changes can't be represented in the manual calculations. Therefore, it is desired to construct means of effective scheduling and updating the parameters continuously.

The organization of the thesis is as follows. The next chapter includes a brief review concerning with job shop scheduling algorithms and data mining techniques. In Chapter 3, we discuss G&T algorithm and clustering techniques. That chapter also includes the issues on how to connect MES database and ERP database. In chapter 4 we discuss the details of integrating G&T algorithm into ERP system. Chapter 5 is last chapter and this chapter summarizes the analysis and results of the thesis. Pathways to future improvement is also presented in this chapter.

CHAPTER 2
LITERATURE REVIEW AND BACKGROUND STUDIES

Scheduling problems need mathematical techniques or heuristic methods in order to find efficient solutions for allocating limited resources onto related task. Increasing customer expectancy and technological improvements have given rise to increase in the difficulty of production systems. Also, firms face stochastic commotions and periodic demands, which results in an unstable application of manufacturing capacity[3].

Many companies can achieve a decrease in inventory levels and increase in productivity and delivery performance with a good scheduling system based on performance criteria determined by the production system.[4]

Thanks to the development of the Industry 4.0 philosophy, companies have begun to have many of their works done on machines or software. The philosophy of Industry 4.0 is based on the fact that jobs are done by machines and manpower is minimized[5]. According to this philosophy, robots and software will take over the place of people. In addition, companies aim to use this philosophy which also includes software automations in decision making processes. This philosophy is very broad-based but to briefly examine the structure on which it is built; Augmented Reality, Internet of things, 3D printing, big data, artificial intelligence can be considered[6]. We can also define the concept of Industry 4.0 as cyber physical systems (CPS). CPS brings together virtual and physical worlds. It is a system in which intelligent objects communicate and interact with each other. When the boundaries between the real and virtual worlds disappear, many innovative applications are expected to be implemented[7]. Today, Industry 4.0 philosophy has not yet been fully introduced into our lives. The main reason of this is economic reasons. Many sensors are needed for instant monitoring of the production flow and integration of decision making systems[3]. On the other hand, companies need to meet some of the requirements of this philosophy for their digital transformation and be prepared for the future.

The biggest advantage of the Industry 4.0 philosophy for the production environment is that instead of the standard business concept, the machines communicate with each other and make decisions on their own according to the changing situations in the production environment. This situation affects working processes. For example, let us consider scheduling functions in the production environment. For scheduling, the machine must be available and the material must be ready to use on shop floor. If the material will be received one day later, this system changes the job on the machine; selects the work which has materials in stock and is suitable for that machine and then restarts the production.

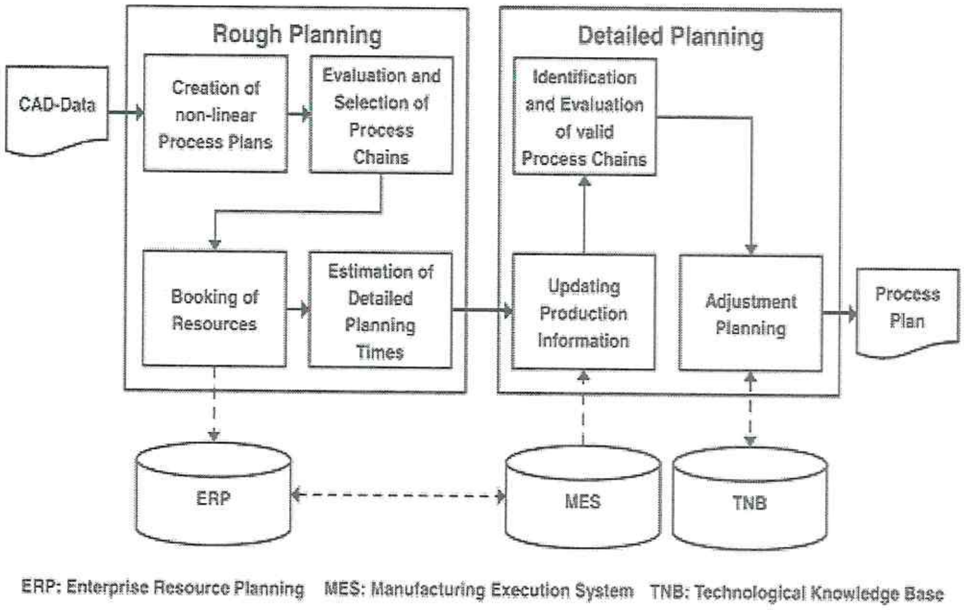


Figure 2.1. Flow Chart Of Adaptive Process Planning

As you see Figure 2.1, the plants firstly implement ERP system (Enterprise Resource Planning) and all of the departments contribute to collect data. Specially these data use in business decisions. The second step is MES (Manufacture Execution System) to collect data area and also give feedback to production area. Industrial revolution- also called Industry 4.0 provides a lot of convenience in data collection and processing. Manufacture Execution Systems (MES) is a dynamic information system application that enables the implementation of production operations, and using existing and accurate data, MES directs and reports facility activities as events.[8].This concept represents a new industry integrating a series of production systems. Convergent technologies that add value to the entire product life cycle. Industry 4.0 was invented in 2011 by universities and private companies by a German initiative of the federal authority. [9].

Industry 4.0 results from advanced production or at the same time Smart Manufacturing approach, an adjustable system that is flexible and sets manufacturing processes for various categories of lines, products and unstable situations automatically. Smart products consider technologies for the product processing, while Smart Products contemplate technologies link to the output offering[10].

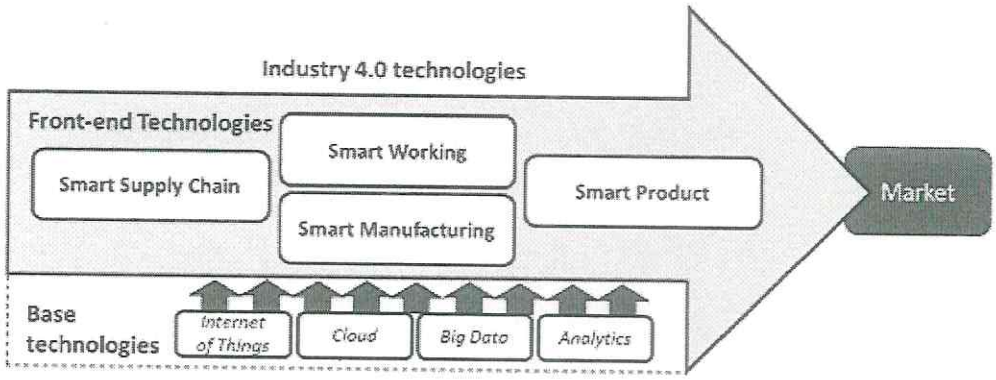


Figure 2.2. Industry 4.0 Technologies

As it is seen in Figure 2.2 industry 4.0 affects all processes of factory. A core element of Industry 4.0 are high-tech physical systems. They can be defined as embedded systems using sensors to collect data and act on physical processes by actuators over digital networks[11].

Technical progress in data storage technology and reduced costs for gathering the data and holding them in stock have remarkably increased both the amount and accessibility of the production data[12]. Analysis of the data cannot often be efficient because of the amount and complexity. Automatic analysis methods are reviewed in the Knowledge Database of Databases in the literature.(KDD)[13].

2.1 Scheduling Problems in General

Scheduling is one of the most important decision making tools which is applied on various industries include manufacturing, purchasing, transportation, logistics and communication. Scheduling is an implementing of limited resources for the best practices via mathematical techniques or heuristic approach to provide conformance to production plan and to be able to catch ODDs (on dock date) of products as promised. As general overview, while the production planning determines the type of products, production time and required quantities, scheduling defines how the production is performed according to loading of current capacity, material availability and inventory levels as well as the sequence of products and determination of resources for specified operations [14].

It can be said that scheduling is looking for the answers of three questions. The first question is the required resources and usage rate of them, production time and sequence of products.

The main reasons of looking for the answers to these questions in the manufacturing environment are to ensure high level of customer satisfaction and to use resources efficiently. In additionally, determination of bottlenecks on manufacturing area is the other important point to ensure efficient usage of resources together with working on cost reduction[15].

These specified purposes can be conflict with each other in practice. For instance, increasing capacity to meet the customer demands will reduce the usage rate of resources.

Thus, providing an efficient production planning and controlling it become a significant advantage in this area. Production planning relates to repetitive activities in production, such as production control, releasing production orders, machine/tool allocation and re-schedule in/out in short-term due to demand changes or unplanned failures on machines, while needed technologies and strategies for production have been determined together with the sequence of manufacturing steps. On the other hand, low-volume production or even one-piece production occur planning inefficiencies. These production types are customer expectations and if company have to meet these expectations in the modern world. In todays world, companies that cannot meet these expectations are difficult to survive in a competitive environment. In this thesis, we will study a scheduling problem that is included in the real life problems cases.

There are many types of scheduling problems each having different characteristics depending on the work flow, machine configurations etc. A general classification of scheduling problems can be seen as in Figure 2.3. [16] [17];[18].

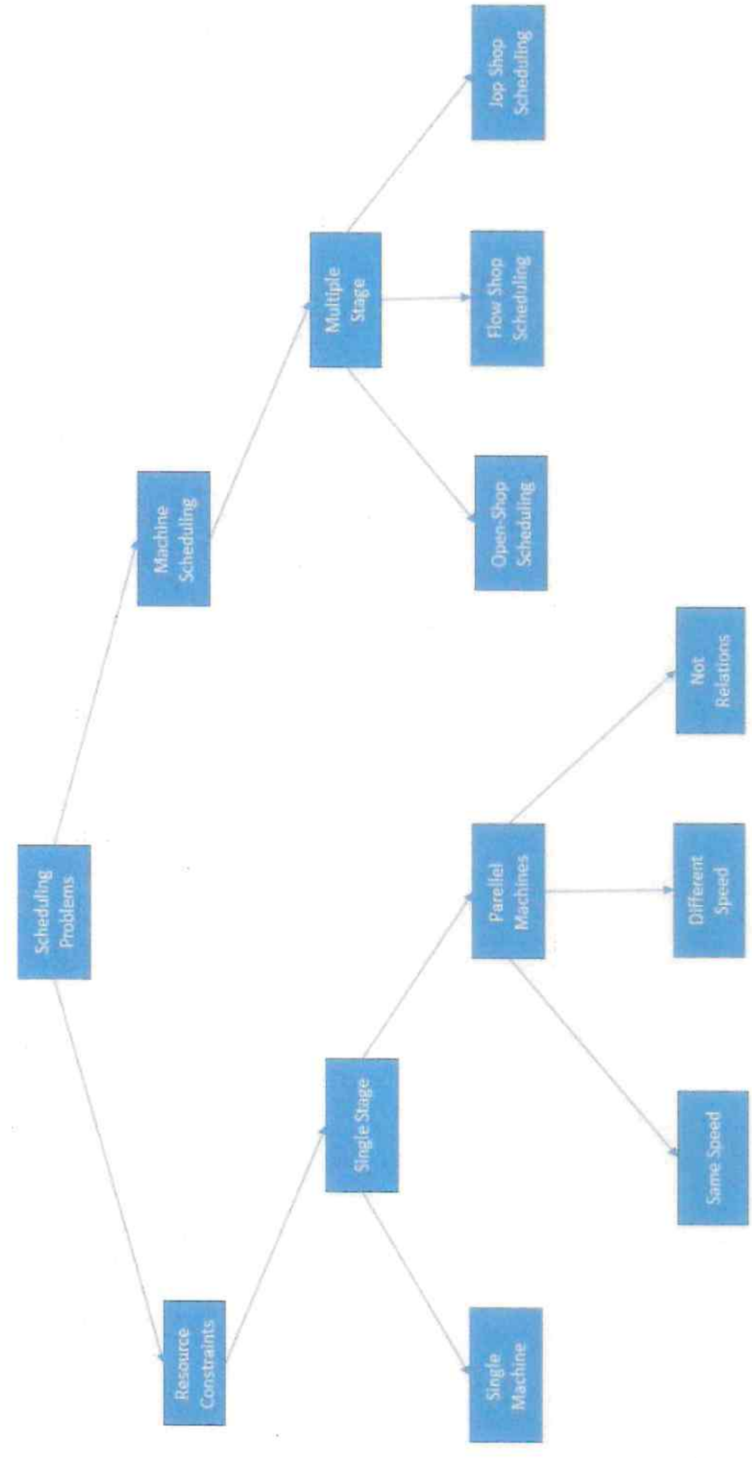


Figure 2.3. Classification of Scheduling Problems

The scheduling problems are stated with three parameters as shown in Figure 2.4. The term " α " is related to the machine environment. The term " β " represents the process characteristics and constraints whereas the term " γ " specifies the objective function. [19]

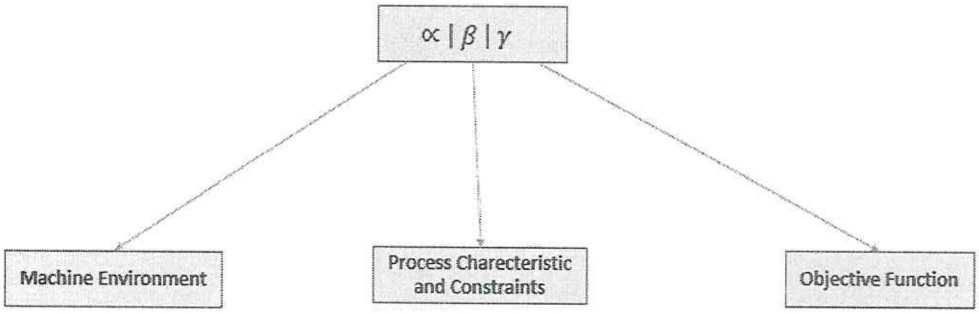


Figure 2.4 Taxonomy of Scheduling

The parameter "α" takes some values as follows:

1. Single Machine (1): A single machine is the easiest possible in the machine environment and a special case of all other mixed machine environments. There is only one machine at floor and jobs have to processed only this machine.

2. Parallel Machines (P_m): There are m identical machines in parallel. The job j requires a single operation and maybe processed on any one of the m machines or any one that belongs to given subset.

3. Parallel Machines at Different Speeds (Q_m): There are m parallel machines at different speeds. The speed of machine i is denoted by v_i. The time p_{ij} that job j spends on machine i is equal to p_j/v_i(assuming job j receives all its processing from machine i). This environment is referred to as uniform machines.

4. Flow Shop (F_m): There are m machine in job floor. Each job can be processed on any m machine. All jobs have same route. In this case some jobs wait for idle machine and this makes a queue. Generally, FIFO (first in first out)

dispatching rule use while one job cannot pass another. If the FIFO rule is effect, then flow shop becomes a permutation flow shop.

5. Flexible Flow Shop (FF_c) : Flexible flow shop is generalization of parallel machine and flow shop environments. There are several identical machines in series at each stage. Each job has to be processed first at stage one, then at stage two and others. Jobs have no priority. The queue at one stage may or may not be processed according to the FIFO discipline.

6. Job Shop (J_m) : In a job shop, each job has a predetermined route to follow. There are variants of job-shop problems in which each job visits the machines at most once or jobs may visit machines more than once.

7. Flexible Job Shop (FJ_c): The flexible job shop is a generalized version of the job shop (J_m) and parallel machines environments. There are many identical work station. Each jobs have its own route to follow through the shop; job j requires processing at each work center on only one machine and any machine can do.

8. Open Job Shop (O_m): There are m machines. Each job must be processed again on each of the m machines. Some of these processing times may be zero. Various jobs can have different routes, no route restrictions.

The parameter "β" takes some values as follows:

1. Release Date (r_j): If this symbol is displayed in the β field, job j cannot start processing before arrival time. If this symbol is not shown in the β field, the process of job j can start at any time.
2. Sequence Dependent Set Up Times(s_{jk}) : s_{jk} specifies sequential preparation times between jobs *j* and *k*. If job *k* is first in the queue, it indicates the preparation time for the job job *k*, and if *j* job is at the end in the queue, s_{jo} is the cleaning time after job *j* (insert and s_j can be zero). If the preparation time between job *j* and job *k* depends on the machine, it contains the subscript i (e.g. s_{ijk}). If there is no sjk in the β field, all preparation times are assumed to be zero or the queue is independent.
3. Preemptions (*prmp*): Jobs don't need to processed at same machine until finish.

A different job can be placed on the machine by stopping the processing of the job at any time. When the interrupted job put back into the machine, it is only processed for the remaining time. This priority rule need, if one job has high priority.

4. Precedence constraints (*prec*); Priority constraints can occur in a single machine or parallel machine environment. One or more jobs have to be completed before the other job is allowed to start. Priority constraints have several exceptions. If there is no *prec* in the β field, jobs are not subject to priority constraints
5. Breakdowns (*brkdwn*); Machine breakdowns indicate that machines are not continuously available. A machine is repaired during an unusable period. If there are a number of parallel identical machines, the number of machines available at any given time is a function of time.
6. Machine eligibility restrictions (M_j); The presence of the M_j symbol in the β field indicates that there are parallel identical machines. The M_j set indicates the machine set that can handle the job j. If the β field does not contain M_j, the job j can be processed on any machine m.
7. Permutation (*prmu*); Flow type refers to queuing discipline according to the principle of FIFO in the job shop environment.
8. Blocking (*block*); This can occur in the job shop. Indicates that when the queue between two consecutive machines has a full or limited capacity, the previous machine cannot send the job that it has finished to the next machine.
9. No-wait (*nwt*): The absence of waiting is a separate situation that occurs in flow-type job shop. Waiting between two consecutive machines is not allowed. For example, it is not allowed to wait in the steel plate rolling mill to prevent drying. In the model with this restriction, tail discipline is FIFO.
10. Recirculation (*rcrc*); It is seen in job shop type or flexible job shop type. Recirculation occurs when a job traverses a machine or workstation more than once.

The parameter "γ" takes values as follows;

1. Makespan (C_{max}); The last process completion time in the system. If this value is small mean that the machine efficiency is high.

2. Maximum Lateness (L_{max}): It is defined as $\max(L_1, \dots, L_n)$. In scheduling, lateness is a measure of a delay in finishing a job on due date, and earliness is a measure of finishing operations before due date. Completion time of job j, C_j; lateness time , L_j; positive lateness, T_j and due date is d_j.

$$L_j = C_j - d_j$$

$$T_j = max\,\{C_j - d_j\} = max\{L_j, 0\}$$

The difference between the tardiness and the lateness lies in the fact that the tardiness is never negative. The *unit penalty (U_j)* of job j is defined as ;

$$U_j = \begin{cases} 1 & C_j > d_j \\ 0 & otherwise \end{cases}$$

3. Total weighted completion time ($\sum w_j C_j$); The sum of the weighted completion times of the *n* jobs gives an indication of the total holding or inventory costs incurred by the schedule. The sum of the completion times is in the literature often referred to as the flow time. The total weighted completion time is then referred to as the weighted flow time.
4. Total weighted tardiness ($\sum w_j T_j$); This is also a more general cost function than the total weighted completion time..

The scheduling problem notations examples are as follows;

$J_m|C_{max}$ denotes a job shop problem with *m* machines. There is no recirculation, so a job visits each machine at most once. The objective is to minimize the makespan. This problem is considered a classic in the scheduling literature and has received an enormous amount of attention. Also in this thesis we study specially on this model.

$1\,|s_{jk}\,|\,C_{max}$: Single machine scheduling problem with sequence dependent set up times and the objective function is to minimize the makespan.

2.1 Job Shop Scheduling Problem (JSP)

A classic Job Shop Scheduling Problem (JSP) has a set of n jobs processed by a set of m machines. Each job is processed on machines in a given order with a given processing time, and each machine can process only one type of operation.

Job Shop scheduling problems are the most complex in terms of classification among the scheduling problems. Because there is no restriction on the number of transactions belonging to a particular one, and there are many production routes to be considered as alternatives. In job shop scheduling, each order has its own process and sequence of operations to be processed on different machines [20]

The solution of a JSP can be depicted in Gantt charts. Figure 2.5 shows a Gantt diagram in which the sequence and timing of the jobs are shown by the machines. [21]. On the other hand, the routes of jobs and their processing times through machines can be tracked on Figure 2.6.

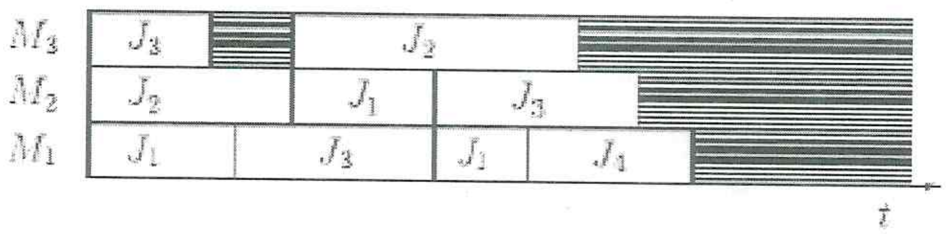


Figure 2.5. Job Based Representation

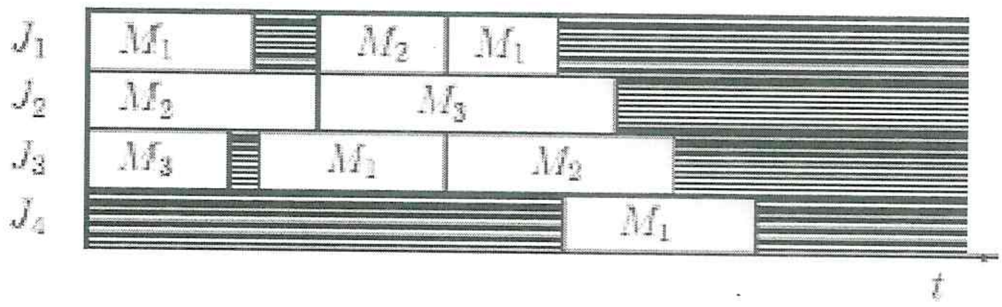


Figure 2.6. Machine Based Representation

There is a 3-job and 3-machine Job Shop scheduling problem example is given in Figure 2.6. Where j is the operation and m is the machine to perform this operation. If we explain the process time, the first job's first operation (J_{11}) is 3 unit. First job's second operation (J_{12}) is also 3 unit. And first job's third operation is (J_{13}) is 2 unit. Also figure 2.6 shows the routes of job. Job one route is machine 1, machine 2 and machine 3. And job 2 route is machine 1, machine 3 and machine 2.

Process Time

Jobs	Operations 1	2	3
j_1	3	2	2
j_2	1	5	3
j_3	3	2	3

Machine Sequence

Jobs	Operations 1	2	3
j_1	m_1	m_2	m_3
j_2	m_1	m_3	m_2
j_3	m_2	m_1	m_3

Figure 2.7. Example of Job Shop Scheduling Problem

In this example we need the notation to show which job operation is assigned which machine. The notation 2.1.1 index means that the 2nd job of the 1st operation is processed on the 1st machine as you see Figure 2.8-a [19, 22]. The solution diagram of the appropriate schedule in the machine and work Gantt diagrams is given in Figure 2.8. The operations of the works throughout the time scale are scheduled on the related machine. As you Figure 2.8-a machine based example, we focus one machine 1. Machine 1 process 2.1.1 then 1.1.1 and 3.2.1. That means machines 1 process second job's first operation and it is 1 unit.1.1.1 notation is machine 1 process first job first operation and it is 3 units. Lastly 3.2.1 is third job second operation on machine one. It is 2 units.

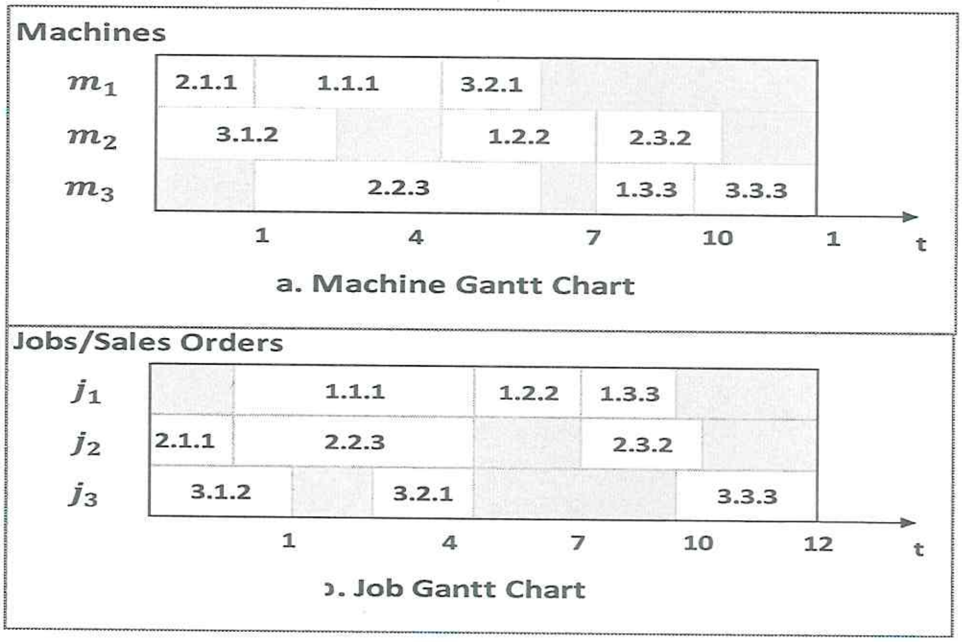


Figure 2.8. Gantt Charts

2.2 General Characteristics of Job Shop Scheduling Problem

The Dynamic Character of Job List: Generally, it is common to assume the list of jobs is known and fixed. In other words, the list is static, and try to solve the problem using the fixed list of jobs. However, the number of works to be processed may change continuously. Therefore, although most of the algorithms for scheduling problems are static, most of the scheduling problems in practice are dynamic.

Number and Specifications of Machines: Generally, all machines are considered to be identical. However, in practice this may not always be the case. The production rate of a particular machine is based on different factors, such as the condition of the machine or the experience of the operator. Depending on the nature of the work and the layout of the job shop, there may be some constraints that prevent the optimization of the objective function.

Number of Proficiency of Employees: The number of employees in the workshop, as well as the variety and number of machines, reveals the capacity of the workshop. Capacity planning is an important aspect of production planning. Many control systems do not include capacity constraints in the problem. Capacity is a dynamic constraint. Failure of a critical machine or loss of a critical worker can result in workshop capacity bottleneck. This is a factor reducing capacity.

Specific workflow: In scheduling problems, there is often a need to process works in a specific order. However, it is not always possible for all works to be available on the machine. Works are formed according to the flow diagram required by the production system and they become ready to be processed. This is due to the fact that the production system and scheduling are independent of each other. Therefore, unfavorable workflows for the algorithm may occur.

Evaluation of alternative rules: Selection of objective function, is the most important factor in the selection of the rule to be used. Sometimes the optimality of more than one objective function may be required. In this case, it becomes impossible to specify a single rule or algorithm. For example, it may be desirable to prevent all works to be processed at the minimum time, as well as to complete any of the works after the deadline.

2.3 Assumptions in Solving Job Shop Scheduling Problems

Generally, the following assumptions are applied for the job shop scheduling problem [20, 23].

a. Each work is complete: although the work consists of different operations, the two operations of the same work cannot be processed simultaneously at all.
b. There is no division: When each operation starts, the other operation must be completed before it can be started on that machine.
c. Each work has "*m*" different operations, one for each machine: the possibility that the work is processed twice on the same machine is not taken into account.
d. No work cancellation: Each work must be processed until it is completed.
e. The time required to move works between machines can be neglected.
f. Work in progress stock is allowed: Next works can wait for the works on the next machine to finish.
g. There is only one of each type of machine: When processing works, it is assumed that there are no more than one machine performing the same work.
h. Machines can be empty.
i. No machine can handle multiple operations at the same time.
j. The machines never fail and are always ready for use during the scheduling period.
k. Technological constraints are known and constant.
l. Randomness is not available. Especially;
    - The number of jobs is known and constant.
    - The number of machines is known and constant.
    - Processing times are known and constant.
    - Preparation times are known and constant.
    - Any qualitative value needed to identify a particular problem is known and constant.

2.4 Performance Criteria in Scheduling Problems

The most common objective in scheduling problems is to minimize makespan. The aim is to finish last operation of the last job as soon as possible. The notation and formulation of JSP is given below with the objective function of minimizing makespan [24]

i: job index$(i = 1, \dots, n)$

j: Index for operations for job (j= $1, \dots, J_i$)

k: machine index$(k = 1, \dots, m)$

J_i: Number of operations require to complete job i

N_k, Set of operations $\{O_{ij}\}$ that can be loaded on machine k

t_{ij}, Processing time of operation O_{ij} in JSP

H, a huge positive integer number

S_{ij}, Start time of operation O_{ij}

C_{ij}, Completion time of operation O_{ij}

O_{ij}, Operation j of job i

w_i, Weight assigned to the objective function i

d_i, Due date of job i

$Y_{iji'j'}$, Decision variable for generating a sequence between the operations O_{ij} And $O_{i'j'}$ in JSP

Objective:

$$Minimize\ [Max\ (C_{1j_1}, C_{2j_{12}} \dots, C_{nj_n})] \quad (1)$$

Subject to;

$$C_{ij} - S_{ij} - t_{ij} = 0 \qquad \forall\ i, j \quad (2)$$

$$C_{i'j'} - C_{ij} + H(1 - Y_{iji'j'}) > t_{i'j'},$$
$$\forall\ (i,j), (i',j') :\ O_{ij} \in N_k, O_{i'j'} \in N_k \quad (3)$$

$$C_{ij} - C_{i'j'} + H(Y_{iji'j'}) > t_{ij},$$
$$\forall\ (i,j), (i',j') :\ O_{ij} \in N_k, O_{i'j'} \in N_k \quad (4)$$

$$S_{ij} \geq 0\ ,\ \forall\ i, j \quad (5)$$

$$S_{ij+1} - C_{ij} \geq 0, \qquad \forall\ i, j = 1, \dots, j_i - 1 \quad (6)$$

$$Y_{iji'j'} = \begin{cases} 1, & if\ operation\ O_{ij}\ precedes\ O_{i'j'} \\ 0, & otherwise \end{cases} \quad (7)$$

Where the start time (ST) is S_{ij} and completion time (CT) is C_{ij} of j_i. We need a large positive number, which is H. In addition, N_k is the set of operation O_{ij} . O_{ij} And $O_{i'j'}$ are the operations which start times are S_{ij} and $Y_{iji'j'}$ is decision variable that generates sequence between these operations.

The constraints (2) are the difference between CT and ST of a transaction is equal to the processing time. This constraint fulfils the assumption that once an operation starts, it cannot be prevented until it is completed. The constraints (3) and (4) provide that two operations cannot be performed on the same machine at the same time. The distinctive constraint (3) be passive when $Y_{iji'j'}$=0 and the distinctive constraint (4) becomes passive when $Y_{iji'j'} = 1$. The constraints set (5) provide that the start time of a process is always positive. The constraint set (6) represents the priority relationship between the various process of a job.

On the other hand, there may be objectives other than minimizing makespan. There may even be multiple performance criteria. For example, the following formulation describes an objective of minimizing the weighted sum of makespan, total tardiness and total idle times of the machines.

$$\text{Minimize } w_1 x \left\{\max[C_{ij_i}]\right\} + w_2 x \sum_{i=1}^{n} \max[0, C_{ij_i} - d_i] + w_3 x \sum_{k=1}^{m} I_k \tag{8}$$

Where

$$\text{Makespan} = \max\left[C_{ij_i}\right], \tag{9}$$

$$\text{Total Tardiness} = \sum_{i=1}^{n} \max[0, C_{ij_i} - d_i] \tag{10}$$

$$\text{Total İdle Time} = \textstyle\sum_{i=1}^{n} I_k \tag{11}$$

There are two new parameters to improve in above equations. One of them is idle time which means to try minimize slack time of machines. And the other one is total tardiness time.

2.5 Job Dispatching Rules in Heuristic Methods

Since JSP is an NP-hard problem and it is not practical to use mathematical model to solve real world scheduling problems, researchers developed heuristic method to solve the problem. Most of the heuristic methods employ some job dispatching rules to obtain initial feasible solutions. The most commonly used dispatching rules for scheduling problems are as follows [19].

First Come First Serve (FCFS): This rule is also known in the literature as first-in-first-out. According to the basic logic of this rule, the works coming to the work centre are taken into the production environment according to their order of arrival and processed in that sequence.

Last Come Last Serve (LCLS): This rule recommends performing a sequence in the work centre in contrast to the order of arrivals. The last work in the work centre is taken first and the second-to-last work is taken in the second place and the order continues in this way.

Do the shortest processing time first (SPT): This rule recommends that works that come to the work centre are sorted by establishing a relationship that is directly proportional to the shortness of the processing time. The output of the production system is maximized and the rate of delayed works decreases with this rule that prioritizes the works with the shortest processing time.

Do the longest processing time first (LPT): According to this rule, the works coming to the work centre are sorted in relation to the length of the processing times. The work that has the maximum processing time among the works that come to the work centre is processed first.

Do the weighted shortest processing time first (WSPT): In scheduling problems, not all works have the same importance level. In this priority rule, in determining the priorities of the works, the processing periods are taken into consideration together with the importance of the relevant works.

Do the work with the maximum number of subsequent processes first (LNS): This rule recommends that works in the production environment are correctly sorted with the number of the remaining processes. According to the rule, the work with maximum number of remaining processes is done first.

Do the work with the earliest deadline first (EDD): According to this rule, which aims to minimize the delays of the works, the work which has the earliest delivery date among the works coming to the system is done first. With the use of this priority rule, the rate of deliveries on the promised dates increases.

Random sorting (SIRO): One of the works coming to the work centre and waiting to be treated is randomly selected. There is no optimization objective.

Do the work with the least slack time first (LSF): Sorting according to this rule is done in direct proportion to the shortness of slack times in the system. By dynamically monitoring the slack times of works, the percentage rates of delayed works are reduced.

Do the work with the shortest preparation time first (SST): This rule recommends that works that are waiting to enter production are ordered in a proportionate to the shortness of preparation times. By giving priority to works with the shortest preparation time, the average delay time can be reduced.

Do the work with the least critical rate first: The critical rate in this rule is calculated by the ratio of the time remaining until the delivery date of work to the production time required to finish that work. According to the rule, works that are waiting to enter production are ranked in direct proportion to their critical rates [24].

Do the work on the critical path first (CP): According to this rule, works are sorted according to priority constraints and the works on the critical path are processed first. The critical path is the way in which any delay will delay the entire project. Therefore, delays in the project are minimized by giving priority to the works on the critical path.

2.6 Data Mining Technique

The amount of information produced and stored on a global scale is unimaginable and is growing more and more today. It defines big data as data sets that are very large and raw, growing in an exponential way, far from being structurally, unresolved even with the help of traditional relational-based database techniques, the majority of which are unstructured and continue to accumulate in an endless manner [25]. Until the 2000s, big data volumes were stored and analyzed. In the 2000s, big data is considered by three components: volume, speed and data diversity [26]. Until the 2000s, the meaning of big data is to collect and analyses large amounts of data in terms of volume. In the 2000s, big data is considered by three components: volume, speed and data diversity [26].

2.6.1 Big Data Definition

In recent years, many researchers have done successful studies on big data. Many articles are published in the literature. For instance, Fortune, Forbes, Business Week, Bloomberg, The Wall Street Journal and The Economist magazines contain important publications.[5] Governments are making major investments in this business. For example, in March 2012, the US Obama Management declared that $ 200M would be invested for big data surveys. An IDC Report forecasts that between 2005 and 2020, the worldwide data volume will increase from 130 Hex bytes to 40,000 Hex bytes and doubled every two years. Some of the social networking sites have between 100-750M users such as Facebook (750M users), Twitter (250M users), LinkedIn (110M users). Big data has aroused great interest in both the industry, research institutions and the government, leading to a new field of research. Mobile Cell Phones, for example, make a great contribution to facilitating our daily lives, obtaining data from people in different ways, and accessing and processing large amounts of data.

The McKinsey Global Institute (MGI) uses the following description for big data: Big Data[27], in view of volume or size, is data sets that exceed the capabilities to acquire, store, analyses and manage data of traditional database software tools. This contingent encryption is subjective and contains a variable description of how big the data should be for a data set to be think about as big data. In this definition, MGI underlines that there is no actual volume threshold for the data to be defined as "big" but that it contingents on the environment. Besides, the description uses the volume of data as a single criterion. As stated herein, this use of the term "big data" may be deceptive since it advises that the concept is primarily related to the data volume. As such, the problem is not a new problem. The basic question should be how to handle the data that is supposed big at a specific point is a long-standing issue in database research.[25] Therefore, considering the waves created by big data, it is necessary to consider dimensions more than volume. Indeed, most publishing expand this description. One of these definitions was indicated in a study by IDC: IDC identifies large data technologies as new generation technologies and architectures, and highlights its economic value by providing high-speed capture to obtain large amounts of data.

Although data volume is a commonly used factor for the quality of Big Data, several important features such as volume, diversity, speed, valuation and accuracy are emerging in the case of Big Data analysis. In the 2000s, big data is characterized by three components: Volume, Speed and data diversity [26]. The English 3V abbreviation (volume, diversity and speed) concept was first used by an analyst Doug

Laney in 2001. This approach provides some basis for quantifying data as Big Data, but does not provide a definitive model since it does not allow assumptions to scale the data. In addition, the value and accuracy of the data are two other commonly used factors along with Gartners 3Vs.

Big data needs new generation databases (called NoSQL databases) that allow you to analyses unstructured data. These new databases are expected to be adopted by companies to meet the growing volume, speed and the need to analyses various data requests. Although big data are common in industry reports, the analysis of actual practices within the firm seems inadequate.

Big Data has three main characteristics: data itself, analysis of data and results presentation. This description is established on the 3V model introduced by Doug Laney in 2001. "Big Data" term was not used by Laney, but Laney expected that the trend in e-commerce would make data management more and more important and hard. Then 3Vs - volume, rate and diversity - was determined as the biggest challenge on data management. When the discussion of the "big data" emerged, writers from the industry and business, in particular, adopted 3V model to identify large data and emphasized that all three should be successful for the solution. However surprisingly, there is no consistent definition in the academic literature. Some researchers use the 3V model slightly. Sam Madden defines big data as too fast, too big or too difficult; where very difficult term is applied for the data that do not appropriate for existing processing tools. Thus, the expression very difficult in this definition is very similar to the data variety. Tim Kraska is moving away from 3V, but he still admits that big data is nothing more than volume. Big data is defined as the data that does not ensure that users receive timely, cost-effective and high-quality responses to data-driven questions, a normal implementation of current technology. Some researches use accuracy 4.V, for example by IBM.

Accuracy means trust in data and is the result of data rate and variety to some extent. The top speed that the data comes from and needs to be processed makes it difficult to continuously clean and re-process to enhance the quality of the data. This situation becomes difficult in deference to diversity. Firstly, it is necessary to perform data cleanup and to provide consistency for unstructured data. Secondly, the diversity of many independent data sources can commonly cause conflicts between each other and causes to become difficult to write metadata for each data item or set of data. Thirdly, it may contain contradictions due to human-generated content and social media analysis, human errors and malicious intent.

2.6.2 Characteristics of Big Data

Big Data does not have a standard definition, but many attempts to identify Big Data can be associated with the following factors[28]:

Volume: The amount of generated and stored data is taken into account. The amount of data determines the value and the potential internal meaning, and ultimately it is evaluated whether the data is actually big data. Data is generated from different sources. According to IDC (International Data Corporation) statistics: data is generated from 44 different sources from sensors to supercomputers, from personal computers to servers, from cars to airplanes. According to IDC statistics, the amount of data will increase from 4.4 trillion gb to 44 trillion gb from 2013 to 2020 and additionally, 20% of digital data was traded in the cloud in 2013 while this ratio will be 40% in 2020. Because Big Data requires scalable storage and requires a distributed query approach, data volume is the hardest part of the work. Big enterprises already have large volume of data that have been collected and stored over the years. These data can be in the form of system logs, documentation, record etc. This data amount can reach a level which traditional database management systems cannot overcome. Data warehouse-based solutions may not be capable of processing and analyzing this data due to the lack of parallel processing structure. Today, very useful information can be obtained from data text, locations, or log files. For instance, e-mail communication standards, user preferences, trends in data-base, and security research are some of the results. A solution is offered by Big Data technologies to create value for this enormous, previously unused and difficult to handle data.

Speed: The speed of incoming and outgoing data is taken into account. It is considered to be the speed that the elapsed time for generating and processing the data meets the expectation. Data flows like floods and should be handled close to real time. The data is constantly flowing to organizations at a great pace. Today, web and mobile know-hows have provided a rapid data flow. Consumer and supplier (provider) interactions have been revolutionized by online shopping, and now, online retailers are using information that keeps customers day-to-day and constantly interacting with them, recommending products and highlighting the organization. Online marketing organizations have many advantages with the ability to obtain instant knowledge. The use of huge amounts of data is becoming increasingly important in the future thanks to the invention of the smartphone.

Data Diversity: Data is produced from different sources in different formats. The data may be in structured or non-structural forms, such as numerical, textual, image, audio, video, etc. Non-structural data constitute 70% to 80% of digital data. 80% -90% of useful information is obtained from non-structural data. Data mining, Natural Language processing etc. branches of science try to interpret this data.

Very few of these data are structured data which are produced by social and digital media. Examples of unstructured data include text documents, video and audio data, images, financial transactions, and communications on social websites.

Traditional databases support the storage of large data, but there are some limitations. These data produced by social and digital media are difficult to comply with traditional regular relational database management structures.

These are not a data that can be easily integrated and required a lot of intervention to be managed by applications. In the meantime, they cause loss of information. Big data believes that even at every bit of data, it can have confidential information. So there is no tolerance for data loss.

One of the big data features is the potential of it to aggregate and integrate various sources on a ground for analysis. There is a rapid increase in public, text-oriented resources with the rise of social media. This is accompanied by an increase in group pages, posts, images and messages in social networking. Mobile phones and GPS can also be considered as another source. For example, companies desire to integrate sensitivity to analysis which come from social media sources with the major data of customer and sales data to amend their marketing needs. The diversity of data, in this case, indicates a general variety of data sources. This intends an increasing volume of various data sources, as well as the structural differences between these resources. This leads to the need to integrate unstructured, semi-structured and structured data at a high level, which is very different from each other. At a lower level, the data sources can still be heterogeneous, even if the resources are structured or semi-structured. The structure or scheme of the two data sources may not be fully compatible.

Data Accuracy: Data accuracy refers to the reliability of data, and it is understood that businesses rely on data to make critical decisions. Accuracy is according to the dictionary means "reality or compliance with reality". However, the term describes the absence of these features in the content of big data. It refers to the confidence that the information given is unclear or false or may be corrupted. There are many various reasons for the unpredictability and inaccuracy of the data. First, if different data sources are brought together, the architecture and structure of the data will probably change. The same attribute name or value may be related to miscellaneous things or attribute names or to the same thing. Therefore, one of IBMs researchers, Je Jones, says, “There is no such thing as a single version of the truth”. In reality, unstructured data is concerned, there is not a scheme yet. If semi-structured data is handled, a scheme can be discussed, as in traditional data warehouse approaches, but it is not more precise and clearly defined. Here the data is simplified in detail, configured, and adhered to a related specification. If the last one - unstructured data- is handled, it is often determined by some possibilities when the information needs to be extracted first, and therefore is not entirely precise. In this sense, the diversity feature of the data directly contradicts the feature of truth. Moreover, even the data of a single source can sometimes be blurred and unreliable. This applies in particular to web resources and human source content. People sometimes do not speak the truth knowingly or leave information. If there is man-work in a process, some errors or inconsistency are always faced.

Data value: Data value is able to generate information from data. In fact, this is the most important component and determines the importance of data. The value here supports decision-making, taking into account the value of the results against efforts to collect, manage, process and analyses large amounts of data. While the other four features have been used to define the basic data itself, the value means understanding the inner face of the data generated during the processing and analysis of the data. Data is typically collected for an instant target. For the purpose of use, data should be evaluated immediately. Of course, although the data value is primarily aimed at the first analysis target, it is allowed to use more than one time. The exact value of the data is closely related to possible future analyses and how data is used over time. It is always possible to reuse, expand the data or combine it with new data.

This is the main reason why data is seen as a more and more valuable asset for organizations. This tendency is to collect and protect the data by thinking that it may have a potential value in the future even if it is not immediately necessary. Another reason for the value of the big data resources is that all data is linked together. When the data sets are combined with other datasets, both datasets provide useful results far beyond the value that is analyzed by itself. In this sense, value can be provided when the data segments of the same or similar group of assets are associated with different data sets[29]. This is called network based value. According to the McKinsey Global Institute, these data have different ways of creating value. For example, large data analysis organizations can produce useful results for process changes or to improve their performance and to better understand the situation. "Big data" clusters can be used and analyzed to customize specific actions, such as customizing actions and market segmentation in marketing. Furthermore, the big data analysis approach can support decision making by pointing out hidden correlations, potential aspects of an action, or some hidden risks. The risk or fraud analysis engines for insurance companies can be given as an example for this. Even in some cases, low-level decision-making can be automated for these engines. Finally, big data can provide new business models, new products and services, or improve existing data. Data on how products or services are used can be used to develop new versions of the product[28]. It can generate real-time status data that leads to completely new services and even business models. The focus of such big data to create the expected value is that it focuses on more complex and deeper analyses. Traditional SQL-based analyses on large data sets have been solved by a data warehouse architecture. However, it should not be overlooked that in the "big data" surveys, more complex analyses can be made with large data sets.

In this sense, big data will be provided by switching to more sophisticated analysis methods compared to simple reports in the traditional data warehouse or OLAP. This will be possible by investigating semi-structured or unstructured data, machine learning, data mining methods, multivariate statistical analysis, scenario analysis and simulation. In addition, large data analysis will enable visualization of the results and thoughts of all or some parts of the data set, thanks to the advanced analysis methods mentioned above[30].

2.6.3 Importance of Big Data

Organizations collect a large number of data every day, but it is important to understand the effects of these data on key performance indicators rather than the large number of collected data. Therefore, when analyzing data from a wide variety of sources, it should be helpful to assist in the following situations: time and cost reduction, customized and optimized market opportunities, new product development, strategy development and smart decision-making. The following results can be achieved in the business environment with big data analysis: Root Reason Analysis can be performed in real time for defects, errors and problems. With Quick Calculations, Risk Portfolio can be made within a few minutes. Precautions can be taken before negative situation occurs in an organization and possible situations can be determined in advance with Fraud Detection and Fraud analysis.

Given the characteristics of the above-mentioned big data, the data is quite big and are produced very quickly. In addition, they do not conform to traditional relational database structures as they are not structured. In this case, an alternative way is required to process this enormous data, which contains a lot of hidden information. Big companies can provide resources to tackle this task, but the amount of data produced every day can easily exceed this provided capacity. The provision of cheaper hardware in this direction enables big data to be processed at a much cheaper cost. It means the ability to quickly analyses big data with a large number of confidential information, customers, market trends, marketing and advertising, equipment monitoring and performance analysis, and much more. So, this is an important reason why many large enterprises need big data analysis tools and technologies.

Big data tools mostly use the in-memory data query policy. Unlike traditional business intelligence (BI) software, which executes queries against data stored on the servers hard drive, big data tools perform queries where data is stored. In-memory data analysis has significantly improved data query performance. Big data analysis not only helps businesses to make better decisions and take a significant place in real-time transactions, but also enables businesses to develop new performance metrics and acquire new revenue sources from the information they obtain. Big Data, beyond object-related databases and relational (RDBMS)

databases, perform better with more complex backups and faster search algorithms.

Using Big Data Technologies may lead to disadvantages such as the inability to protect the confidentiality of the data as well as some benefits. In terms of privacy, some companies sell customer data to other companies, which is a major problem. In real-time data flow samples, high-speed data restricts processing algorithms spatially and temporally. Therefore, certain requests must be met to process such data. A new technological infrastructure should be developed for the common functionality to process and analyses the different types of Big Data produced by the services with the gradual increase in the amount of data. A large amount of data types should be analyzed by facilitating fast and effective decision making with this infrastructure.

2.6.4 Big Data Analysis

Data analysis primarily requires that large amounts of data that could affect an enterprise be considered and examined[31]. However, the complexity of the data that needs to be analyzed and the necessity of special algorithms that support such operations for various applications are things that complicate analysis. Data analysis has two main objectives: to understand the relationships between different characteristics and to develop effective data mining methods that can accurately predict future observations.

Nowadays, there is a great need for techniques that can rapidly analyses such a large amount of data. Among the analysis techniques that can be used are data mining, visualization, statistical analysis and machine learning [32]. For example, data mining can automatically discover useful and interesting patterns in a large data set. In this respect, data mining is widely used in various fields such as basic sciences, engineering, medicine and trade. Taking advantage of this technique, in the business world, very useful confidential information is revealed from a large amount of data [33].

Most of the techniques related to large data analysis based on data mining and statistical analysis have been developed and based on relational database technology, data warehouse and ETL.

Big data analysis allows companies to analyses structured, semi-structured and unstructured data. The Mckinsey Global Institute published a study on Big Data in June 2011. The study is divided into two groups as Enterprise Analysis and Academic Research Analyses: In corporate analysis, analysis teams use their expertise in statistics and data mining. In the Academic Analysis, the researchers analyses the data to test hypotheses and build theories.

In the Big Data analysis, the researchers found that the data produced was divided into various large data applications as follows:

Structural Analysis: In structural analysis, large amounts of data are generated in the fields of business and scientific research. These data are data from various research areas such as E-commerce and managed by databases, data warehousing and OLAP tools.

Text Analysis: Text used in text analysis is one of the most common forms of storing information, and it is the data from many different areas such as E-mail communication, documents, advertisements and social media content. Text Analysis refers to the process of extracting useful information from large texts. Text mining is based on text representation and natural language processing.

Web Analysis: The purpose of web analysis is to extract information from web pages. Web analysis is also called Web mining.

Multimedia Analysis: Multimedia analysis refers to interesting information and semantic terms drawn from multimedia data. It covers many topics such as Audio Summarization, Multimedia description, Multimedia indexing and retrieval.

Mobile Analysis: Mobile analysis includes devices such as RFID, mobile phones, Sensors, etc. Mobile data traffic is increasing at a great pace since the end of 2012. The increase in the volume of application data in this area leads researchers to mobile analysis.

2.6.5 Big Data Analysis Techniques

Large data analysis can be applied to custom data types. Besides, many traditional techniques related to statistics and computer science for Data Analysis can still be widely used in large data analysis [13].

Data mining: One of the most important terms related to data-based decision making and is a data analysis tool used to reveal interesting previously unknown patterns. In data mining, information that is hidden from missing, fuzzy, and noisy data, but potentially great value is extracted. Ten of the most common data mining techniques are described in the IEEE International Data Mining Conference. These algorithms include classification, regression, clustering, association analysis, statistical learning and link mining in the investigation of research problems in Big Data.

Association rule: Includes a set of techniques to explore interesting relationships between variables in large databases[13].

Addition Source: The aggregation source typically collects data from a large number of people through an open search through the Web2.0 tool. This tool is used to collect data rather than analyzing it.

Text analysis: Most of the data is in the form of text. Text analysis is the process of converting unstructured text data into meaningful data.

Cluster Analysis: Cluster analysis is a data mining technique which divide big groups to small groups that similarities are unknown. Clustering analysis groups, the objects statistically according to specific rules and characteristics[34]. It distinguishes objects with specific properties and distributes them accordingly. Objects in the same group are very similar, and groups are very different from each other.

Statistical Analysis: In statistical theory, uncertainty and randomize are modelled according to probability theory. Inferential and descriptive anal can be made from large data via statistical analysis. Inferential statistical analysis can formulate inferences about the data subject and random variations, while descriptive statistical analysis can define and summarize data sets. Generally, statistical analysis is widely used in the fields of medicine and economics.

Regression Analysis: Regression analysis reveals the correlation between variables. These variables are dependent and independent variables. It defines the dependent relationships between variables on the basis of experiments or observations. By using regression analysis, complex and uncertain correlations between variables can be simplified and edited.

CHAPTER 3
METHODOLOGY

Although various formulations of mathematical models have been proposed for job-shop scheduling problems, it is not practical to use them to solve real life problems. Because they take too much time. For this reason, many heuristic methods have been developed to overcome this refractoriness and find near optimal solutions in relatively shorter times. One of those algorithms is the Giffler and Thompson (G&T) algorithm [2].

3.1 Giffler and Thompson Algorithm

The heuristic methods usually take an advantage of a rule as priority or dispatching. It means that, a "rule" is used for selection of operation on a machine from a list including all operations which are unscheduled. These are some of common Priority-rules:

1- Shortest Processing Time First (SPT),
2- Earliest Due Date (EDD),
3- Least Slack First (LST),
4- Critical Ratio (CR).

These rules deliver first initial feasible solutions. Then some improvement algorithms such as Simulated Annealing, Tabu-Search and Genetic Algorithms are employed to improve the quality of the solution. In this thesis a G&T algorithm is executed for the job-shop scheduling problem[35].

Classic job-shop model has been studied in our thesis. It is considered that N jobs $\{J_1, J_2, \ldots, J_N\}$ to be processed on M different machines $\{M_1, M_2, \ldots, M_M\}$. The notation o_{ij} is operation j of job i. We defined the production order belongs to the job on related machine previously. Every production order o_{ij} is progressed in the required machine by the fixed processing time t_{ij}.

The JSP aims to specify process sequences in accordance with the strict necessaries and to optimize a performance measurement. In many cases, the makespan, which means the needed time to complete entire tasks, is selected as a performance measurement.

The schedules generated in real life can be divided into three classes as follows.

Semi-active schedules: If the machine sequence is not changed, no operation can start earlier.

Active schedules: There is no any operation is able to start earlier without delaying other operation or without breaking the priority restrictions.

Non-delay schedules: There is no machine can be left idle if there is another operation is able to start on that machine.

The connection of various schedule types is shown on Figure 3.1. It can be seen on this Figure which all non-delay programs are active programs, and that all active programs are semi-active programs. Optimal schedules (to keep the makespan at minimum) are always active schedules [34]. This observation leads to a significant decrease in the research area of high value due to the NP hardness of the JSP [36].

JSP is NP-hardness problem and this type of problems need to time for optimal solution. 10 jobs in 10 machine problem [2] formulated at 1967 (Fischer and Thompson) but there is no solution yet. However, in real life scheduling process are very fast. Generally, in real life schedule of jobs change several times in one day. So nobody waits for the optimal sequences of jobs on machines[37]. So we use some dispatching rules (Earlies Due Date, Least Set Up Time, Critical Ratio) to find near optimal solution to use in production environment. This rules also need one algorithm and in this study we use Giffler and Thompson Algorithm with these dispatching rules. [38]

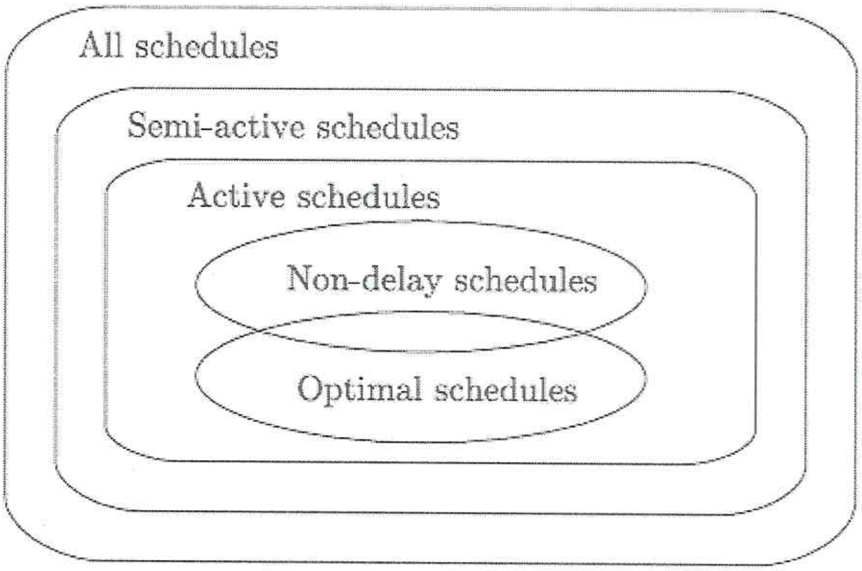


Figure 3.1. Types of Schedules

The presentation of the Giffler-Thompson algorithm [27] uses the following symbols:

- o_k is the operation k $(1 \leq k \leq nm)$.
- p_k is the operation time of operation o_k.
- P_t is the partial schedule of scheduled operations.
- $(t - 1)$ is scheduled operations.
- S_t is the set of operations. Those operations are placed into the schedule at iteration t. They must be run before the other operations in S_t,
- Operation in S_t, are placed in the schedule P_t.
- ρ_k is earliest time.
- Operation o_k from S_t can start at ρ_k .
- b_k is earliest time to terminate the operation $o_k$from S_t ,

$b_k = p_k + \rho_k$.

The Giffler -Thompson algorithm continues as follows by defined symbols:

1. Step : Let $t = 1$ and $P_1 = \{\}$. $S1$ is operations' set without previous ones.

2. Step : Find $b^* = min_{o_k \in S_k}\{b_k\}$ and the machine M^* at which o_k with b^* is completed. Select optionally within the multiple M^*.

3. Step : Select an operation o_j from S_t such as:

- operation o_j requires machine M^*, and
- $\rho_k < b_k$

4. Step: Proceed to next iteration:

- add o_j to P_t, result in P_{t+1}
- remove o_j from S_t and create S_{t+1} by adding the successor of o_j to S_{t+1} (unless o_j is the last operation)
- $t = t + 1$

5. Step : If $S_t \neq \{\}$ go to 2. step. Else Stop.

At each iteration, a conflict set S_t is created by the G&T procedure from chosen & scheduled operation. Until all operations are scheduled, the recurrent G&T procedure should be repeated. Feasible active schedules can be created by this algorithm and the order of priority is kept same.

Stages 2 and 3 ensure that no process is allowed to start earlier, regarding to the priority limitations resulting in an active plan, without delaying another process in the G&T

algorithm. One can create all active programs and the corresponding search field, taking into account all operations in S_t

By changing steps 2 and 3 (see below), non-delayed programs can also be created, such that a machine never idle when it is available to run.

Step 2 Find $\rho^* = min_{o_k \in S_k}\{\rho_k\}$ and the machine M^* at which o_k with ρ^* is carried out. If there are multiple M^* , select optionally.

Step 3 Select an operation o_j from S_t such that:

- this operation o_j requires machine M^*,
- $p_j = \rho^*$

To shorten the search area, the non-delay (ND) variation can be used further which is one of the G&T algorithms [2].

However, the reduced search area is endangered by risk that the optimal solution is not taken into account. [39] refers to find that even an active scheduler can be more convenient in other cases, a non-delay scheduler makes the quality of the solution better in some cases. In terms of genetic algorithms, we can use G&T algorithm and its non-delay in order to turn chromosomes into workable schedules [36]. Crossover or mutation might outcome non-compliance with the order of priority on chromosomes.

Incorrect chromosomes are forced by the G&T algorithm to create a correct/feasible schedule.

Other genetic algorithms use the G&T or ND algorithm to convert a chromosome into an applicable active or non-delay schedule with indirect representation [24].

3.2 Data Mining Tools

G&T algorithms will be used as scheduling algorithm. Scheduling is performed by these algorithms under certain assumptions and in consideration of certain data. But in real life, production lines show continuous fluctuation. If we consider a machine where you make each set up in 180 minutes last year, it is improved to 120 minutes after certain investments. There is a similar situation for cycle times that are constantly changing with the improvements made in the production line each month. As a result, these parameters need to be updated continuously before the algorithm works. However, in order to recognize the connections between set up and cycle times or breakdown times, a complementary data analysis technique was required in the study. Data mining includes all of the techniques used to make meaningful and concise information for us in the big data. Recently, scheduling algorithms provide more optimal results, especially with the use of machine learning techniques. We plan to use clustering method from data mining techniques.

Clustering use to explain classification of individual or objects.[40]. For this purpose, the assets in the example taken divided into groups according to the profile of individuals. In other words, the purpose of clustering is to beings (individuals or objects) small create custom groups, then enter to reveal the profile of assets. [32] Another target is to group similar elements and reduction to group number. Clustering methods are hierarchical and non- hierarchical but some research shows us that the methods more than two. Hierarchical methods use dendogram (tree map) to understand easier. Mostly use in non-hierarchical clustering methods are k-means and maximum likelihood technique. In this section, after a brief look at hierarchical clustering methods, we review in detail k-means clustering.[41] Cluster analysis encounters many of the same issues that we have discussed in the chapters on classification

How is similarity measured?

Basically, we focus on Euclidean distance between records:

$$d_{euclidean}(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Where $x = x_1, x_2, \dots, x_m$ and $y = y_1, y_2, \dots, y_m$ represent the m attribute values of two records. Certainly, there are many other measurements, such as city-block distance:

$$d_{cityblock}(x,y) = \sum_i |x_i - y_i|$$

or Minkowski distance. A general exponent q represents the general state of the two metrics above :

$$d_{Minkowski}(x,y) = \sum_{i} |x_i - y_i|^q$$

We can redefine the "different from" function to compare the i^{th} attribute values of a record pair for categorical variables:

$$different\ (x_i - y_i) = \begin{cases} 0 & if\ x_i = y_i \\ 1 & otherwise \end{cases}$$

Here, x_i and y_i are called as categorical values. Then, we can use different (x_i, y_i) for the term i^{th} in the above Euclidean distance metric.

Classification algorithms, clustering algorithms and normalizing are to be used for best performance. Thus, no variable or its subset has any effect on the analysis. Minimum or maximum normalization or Z-score standardization can be used by analysts as they are discussed in the previous sections:

$$Min - \max\ normalization\colon\ X^* = \frac{X - \min(X)}{Range(X)}$$

$$Z - score\ standardization\colon X^* = \frac{X - \text{mean}(X)}{SD(X)}$$

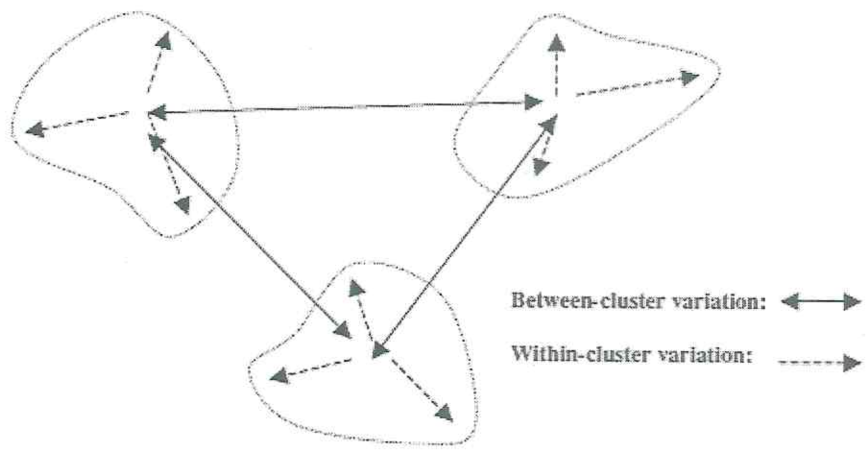


Figure 3.2. Variations and Clusters

All clustering methods have the purpose of defining registration groups such that the similarity within one group is very high and that of records in other groups is very low.

In other words, as shown in Figure 3.2, the clustering algorithms aim to create record sets as an example that the between-cluster variation (BCV) is greater than the within-cluster variation (WCV). This difference is much alike the concept behind the variance analysis.

3.3 K-Means Clustering

K-means clustering algorithm is one of the most used algorithms in data mining. Clustering algorithms are algorithms that automatically divide data into smaller clusters or subsets. Algorithm puts statistically similar records into the same group. An element is only allowed to belong to a set[28]. The cluster is the value that represents the central cluster. The letter of k in the name of the algorithm actually indicates the number of sets. The algorithm also looks for the number of "k" clusters that will minimize the Carcass Error Function commonly used in error calculation. The given number of data sets "n" is placed in k sets in such a way that this error function is minimized. Therefore, cluster similarity is measured by the proximity of the values in the cluster to the mean. This is the center of gravity of the cluster [40]. The value in the center of the cluster is the representative value of the cluster and is called the medoid.

There are two important goals:

- The values in the set should be most similar to each other,
- Clusters should not be as alike as possible.

To perform these requests, the following steps must be performed on the algorithm side, respectively:

- Determine the center of classes,
- Classification of samples by distance,
- Determine the new center of classes after classification,
- Repeat steps 2 and 3 algorithmically until desired.

Let $C_1, C_2, \dots, C_K$ indicate sets containing the indices of the observations in each cluster. These properties are:

1. $C_1 \cup C_2 \cup, \dots, \cup C_K = \{1, \dots, n\}$. This means, that is the part of the at least one of the K sets.

2. $C_1 \cap C_{k'} = \emptyset$ for all $k \neq k'$. Cluster do not overlap : no observations are part of more than one cluster.

Let's think that i[th] observation is in the kth cluster, then $i \in C_k$. The within-cluster variation for cluster $C_k$is a measure $\mathrm{W}(C_k)$ of the amount by which the observations within a cluster differ from each other[13]. Thus, the problem is below.

$$minimize_{C_1,\dots,C_K}\{\sum_{k=1}^{K} W(C_k)\} \qquad 1$$

To solve the equation (1) is right way but to make it possible we need to define intra-cluster diversity. The most popular concept is Euclidean distance. That is, we define

$$W(C_k) = \frac{1}{|C_k|}\sum_{i,i' \in C_k}\sum_{j=1}^{p}(x_{ij} - x_{i'j})^2 \qquad 2$$

where $|C_k|$ denotes the number of observations in the kth cluster. In other words, the variation within the cluster is the sum of the double-square Euclidean distances between observations in the cluster set divided by the total number of observations in the cluster set. Equation (1) and (2) are optimization problem that defines K-means clustering,

$$minimize_{C1,\dots,C_K}\left\{\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{i,i' \in C_K}\sum_{j=1}^{p}(x_{ij} - x_{i'j'})^2\right\} \qquad 3$$

Now, we want to find an algorithm to solve (3) - the method of (3) separating observations into K clusters will be minimized. In fact, this is a very difficult problem to solve precisely because there are almost *Kn* ways to divide observations into K clusters. This is a big number unless K and n are small! Fortunately, a very simple algorithm can be shown to provide a local optimum - a pretty good solution - to the K-tool optimization problem (3). This approach is described in Algorithm 3.1.

Algorithm 3.1 K-Means Clustering

1. Randomly assign a number, from 1 to K, to each other of the observations, these serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   a. For each of the K clusters, compute the cluster centroid.
   b. Assign each observation to the cluster whose centroid is the closest (where closest is defined using Euclidean distance).

Algorithm 3.1 is guaranteed to decrease the value of the objective (3) at each step. To understand why, the following identity is illuminating:

$$\frac{1}{C_k}\sum_{i,i'\in C_K}\sum_{j=1}^{p}(x_{ij}-x_{i'j'})^2=\sum_{i\in C_K}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2 \qquad 4$$

Where $\bar{x}_{kj}=\frac{1}{C_k}\sum_{i,\in C_K} x_{ij}$ is the mean for feature j in cluster C_k. n step 2 (a), each step exists to minimize frame deviations. In step 2 (b), the algorithm exists for continuous operation and development (4). Considering that the result of this algorithm will no longer change. Ultimately this equation will give an optimal result.

CHAPTER 4
EXPERIMENTS AND RESULTS

4.1 Experimental Set up

Packaging production is one of the hardest processes to meet deadlines with nearly 200 minutes set up times and variable customer expectations. Especially in the production of paper bags set up time is nearly twenty hours. When each type of bag is produced, the planning department tries to continue the same production to reduce set up time. This causes three major disadvantages.

a) High product stock level and Raw material stock level

b) Production of materials that have not arrived at the deadline yet, inability to produce products which have close deadlines and risk of missing the deadlines

c) Quality problems caused by trying to be quick for on time production.

In order to eliminate these disadvantages, various studies have been tried in the company, but the studies have failed at a certain point. When we look at the basic points of failures, there are a few factors have been seen such as not receiving any system support and computer-aided scheduling, not making inventory management effective before scheduling, and non-balanced use of manpower.

There is an active ERP (Enterprise Resource Planning) system used within the company and MES (Manufacture Execution System) software is also invested this year.

All of the company process manage in ERP system. Sales order to deliver products process follow up in ERP system. Company insert sales order in ERP system and run MRP (Material Resource Planning). MRP explore BOM (Bill of materials) and create a purchase order for raw material. Inventory manage also in ERP System. Also company use ERP system for insert production information data. And lastly deliver process and invoice process follow up in ERP system. It means all department connect ERP system, consequently all departments connect each other with ERP system.ERP software is considered the nervous system of companies. Therefore, the role of these software's in increasing productivity within the company has great importance.

MES software used in production environments started to directly affect scheduling algorithms. In this model, Giffler and Thompson which is one of the job shop type scheduling algorithms, has been studied according to the best set up time and prioritization constraints. Additionally, at this stage, the shortest processing time and shortest delivery time were tested by used different algorithms on C_{max} and the best algorithm was selected.

4.1.1 Packaging Industry Process Standards and Parameters

We can explain the route of the products in the packaging sector as follows;

Printing Process: The printing process is the transfer of the ink onto the film or paper according to desired customer design. Here the customer design is engraved on copperplates or rollers. It is ensured that this design leaches into the film according to specified criteria's. There are two printing machines within the company that can print 300 m film or paper per minute.

Lamination Process: Lamination process is a process that allows a second layer to be directly applied to raw films that leave the printing process or do not undergo the printing process. The objective of this process is to add a further layer onto the printing film to make the product conform to customer standards.

Control Process: It is a special process to review the whole printing process in products where quality control catches nonconformity. All printed material is connected to the control machine. It is possible to find out and mark which meters of the print is incorrect thanks to cameras. By removing the faulty distance, the remaining part is reassembled and made ready for the next process.

Slitting Process: Slitting process is the process where all the bobbins used in production are sized. For instance, if the customer requests packages in 20 cm wide, the whole bobbin (100cm) is put into the slitting process and 5 pieces' bobbins in 20 cm wide have been produced and delivered to the customer.

Bag Process: It is the process in which printed, unprinted, laminated and non-laminated paper bobbins are sized according to customer requirements and transformed into carrying bags or packages. This process has become the strongest production line of sponsor company in recent years and produces lots of bags in various sizes and specifications.

Table 4.1. Production Process Flow-Chart

Incoming	Printing Process	Lamination Process	Slitting Process	Bag Process	Product
Paper Bobbin + Lamination Material + Lamination Glue + Paper Glue + Ink	X	X	X	X	Printed Laminated Paper Bag
Paper + Lamination Material + Lamination Glue + Paper Glue		X	X	X	Unprinted Paper Bag
Film + Ink + Lamination Material + Lamination Glue	X	X	X		Printed Film
Film + Lamination Material + Lamination Glue		X	X		Unprinted Laminated Film
Paper + Paper Glue				X	Unprinted Un-laminated Paper Bag

The matrix of set up times between all production types within the company was made primarily to ensure that the Job Shop Scheduling algorithm can work with the best set up and shortest delivery time algorithm. The parameters of set up times for production are as follows;

Material Change

Glue Change

Mould Change

Bellows/Tube Change

Length Change

Number of Floors

4.1.2 Packaging Industry Process Standards and Parameters

Table 4.2: Work Centre Based Set up Times Matrix

Work Centre	Work Centre Definition	Details
1108	Printing Process	Printing --> Printing shape
1108	Printing Process	Printing --> Material Change
1108	Printing Process	Printing --> Mat Lak
1108	Printing Process	Printing --> Ink Change
1108	Printing Process	Printing --> Cylinder Width
1109	Printing Process	Printing --> Printing shape

1109	Printing Process	Printing --> Engraving Change (Various)
1109	Printing Process	Printing --> Material Change
1109	Printing Process	Printing --> Roller Circumference
1109	Printing Process	Printing --> Ink Change
1110	Printing Process	Printing --> Printing shape
1110	Printing Process	Printing --> Engraving Change (Various)
1110	Printing Process	Printing --> Material Change
1110	Printing Process	Printing --> Roller Circumference
1110	Printing Process	Printing --> Ink Change
%	Slitting Process	Slitting --> Knife Type
%	Slitting Process	Slitting --> Bobbin Slitting Width
%	Slitting Process	Slitting --> Belly Diameter
%	Slitting Process	Slitting --> Material
%	Slitting Process	Slitting --> Winding Direction
%	Slitting Process	Slitting --> Number of Sets
%	Lamination Process	Lamination --> Width Change
%	Lamination Process	Lamination --> Material
%	Lamination Process	Lamination --> Glue Type
%	Slitting Process	Bag --> Mould
1401	Bag Process	Bag --> Length End
1401	Bag Process	Bag --> Bellows/Tube
1402	Bag Process	Bag --> Length End
1402	Bag Process	Bag --> Bellows/Tube
1403	Bag Process	Bag --> Length End
1403	Bag Process	Bag --> Bellows/Tube
1404	Bag Process	Bag --> Length End
1404	Bag Process	Bag --> Bellows/Tube
1405	Bag Process	Bag --> Length End
1405	Bag Process	Bag --> Bellows/Tube
1406	Bag Process	Bag --> Length End
1406	Bag Process	Bag --> Bellows/Tube
1407	Bag Process	Bag --> Length End
1407	Bag Process	Bag --> Bellows/Tube
1409	Bag Process	Bag --> Length End
1409	Bag Process	Bag --> Bellows/Tube

As you can see Table 4.2, the preparation matrix table above, the standards required for each job to move to another job have been set. Printing process is done on 1108 work centre. There are two types of printing. Reverse printing and flat printing. During the production of the product with reverse printing or flat printing, the set up time will occur for each printing type transition and the printing machine will be provided to switch between printing types.

For each of these cases, summary definitions are made in the system. It is expected that the Giffler and Thompson algorithm will work with minimum set up and tie breaking rule. The set up time for the printing operation, which will result from the change of printing shape, is given below. This definition is based on data collected on the shop floor for each set up type. The following code structure is made via CANIAS and developed with its own TROIA language. As you see the Figure 4.1, to describe this coding briefly, if the print pattern changes, accept set up time as 180 min. If the equipment used in printing changes with the type of printing, 5 min set up time will be added for each plate.

The related data are defined separately for all types of preparation.

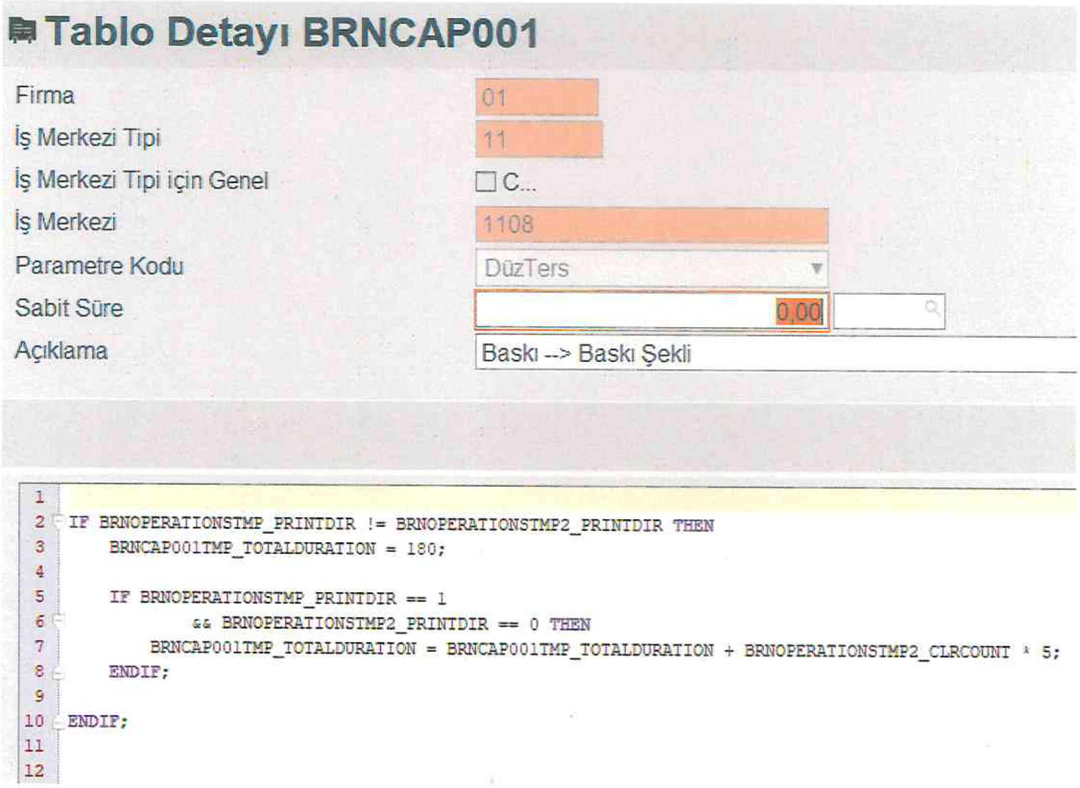


Figure 4.1. Set Up Time System Design Code

Firstly, a simple modelling can be used to explain this setting table and usage of it. In an any machine that produces 4 kinds of jobs in a single work centre, with the assumption that there are productions of type a, b, c, d and also 24 different sorting alternatives as below;

(a-b-c-d), (a-b-d-c), (a-c-b-d), (a-c-d-b), (a-d-b-c), (a-d-c-b), (b-a-d-c), (b-a-c-d), (b-d-a-c), (b-d-c-a), (b-c-a-d), (b-c-d-a), (c-a-b-d), (c-a-d-b), (c-b-a-d), (c-b-d-a), (c-d-a-b), (c-d-b-a), (d-a-b-c), (d-a-c-b), (d-b-a-c), (d-b-c-a), (d-c-a-b), (d-c-b-a)

So, 4! = 4.3.2.1 = 24 alternatives are obtained. One of these rankings is optimal for Set up.

In case the work types are 10;

10! = 10.9.8.7.6.5.4.3.2.1 = 3.628.800 different sorting. During the preparation of the document, 1109 machines had 42 different operations. 42!

Continuing on the same example, it is necessary to know how much set up is done to switch from each of the a-b-c-d jobs to find out which of the 24 alternatives creates at least set up time, Figure 4.3 is an example for this situation.

This means that 4x4 = 16 different set up times must be registered. This data is a matrix structure as follows. When we have this data, it is possible to find out which of the 24 alternatives provides the best set up time.

Table 4.3: Set up Example

Type	a	b	c	d
a	0	10	20	30
b	20	0	40	50
c	10	30	0	60
d	50	45	34	0

The table in which data for this matrix structure is held is IASCAP002 on CANIAS. For example, for a cell coloured in red on the matrix, data such as the following should be registered in IASCAP002.

Figure 4.2. Set-up Times Dialog Box in Canias

Here, work "a" refers to the printing operation of the "carrying bag a" product information form, and work "b" refers to the printing operation of the "carrying bag b" product information form.

This means that jobs and certain types of set up must also be registered in the IASROU004 setting table. However, it should be noted that the types of preparation on the basis of each operation of each product data sheet should be recorded in the IASROU004 table, not on the basis of the product data sheet.

In CANIAS standard, the set up types in IASROU004 are selected on the routes of the works. In IASCAP002, the transition times of these set up types are recorded, and Set up optimization works on this data. In sponsor company, product trees and routes are managed via PIF.

Suppose only 1109 machines have 200 different jobs. This means that 200 different set up types must be opened in IASROU004, they must be defined in the PIFs for printing operations and the calculation of 200 x 200 = 40,000 data must be entered into the IASCAP002 table. Since the production structure is specific to the order, when a new type of work comes for this machine, the new set up type is required to be entered by calculating the new 401 number of 201 x 201 = 40,401 required data after it is opened in IASROU004. This is no longer manageable when all types of jobs and operations are considered.

A BRNCAP001 support table has been developed for company to manage this situation. This table is based on the preparation of the data needed by the set up optimization algorithm used by the system, obtained during the scheduling of IASCAP002 data.

In the case of the 42 jobs mentioned above, firstly there should be a set up type for each of these jobs. Afterwards, these types should have transition times.

During the study, the set up types which codes have the first 4 characters of the work centre were given to the jobs on the work centre. The set up types 11090001, 11090002, 11090042 are assigned to all operations listed in CAPT01. Then the transition time of 42 x 42 = 1764 jobs from each other is calculated according to the formulas in BRNCAP001.

4.1.3 Production Line Capacity Settings

The formulas of the parameters that determine the set up duration according to each work centre or work centre type are entered to BRNCAP001 to calculate the transition times from work to work.

Continuing from the example of 1109, it is stated that 5 different parameters affect the set up time during the transition from work to work.


BRNCAP001

Firma: 01 | İş Merkezi Tipi: 11

	FK	İş Merkezi ...	Genel?	İş Merkezi: ~(3)	Açıklama
11	01	11	☐	1109	Baskı --> Baskı Şekli
12	01	11	☐	1109	Baskı --> Klişe Değişimi (Çeşit)
13	01	11	☐	1109	Baskı --> Malzeme Değişimi
14	01	11	☐	1109	Baskı --> Merdane Çevresi
15	01	11	☐	1109	Baskı --> Mürekkep Değişimi

Figure 4.3. Display Setting Table

In the printing and lamination processes, the highest value calculated from the formulas given in BRNCAP001 is taken as set up, and the sum of the calculations made on the formulas given in the slitting and bag is taken as set up. For the 1109 work centre, review of the Print Type parameter;

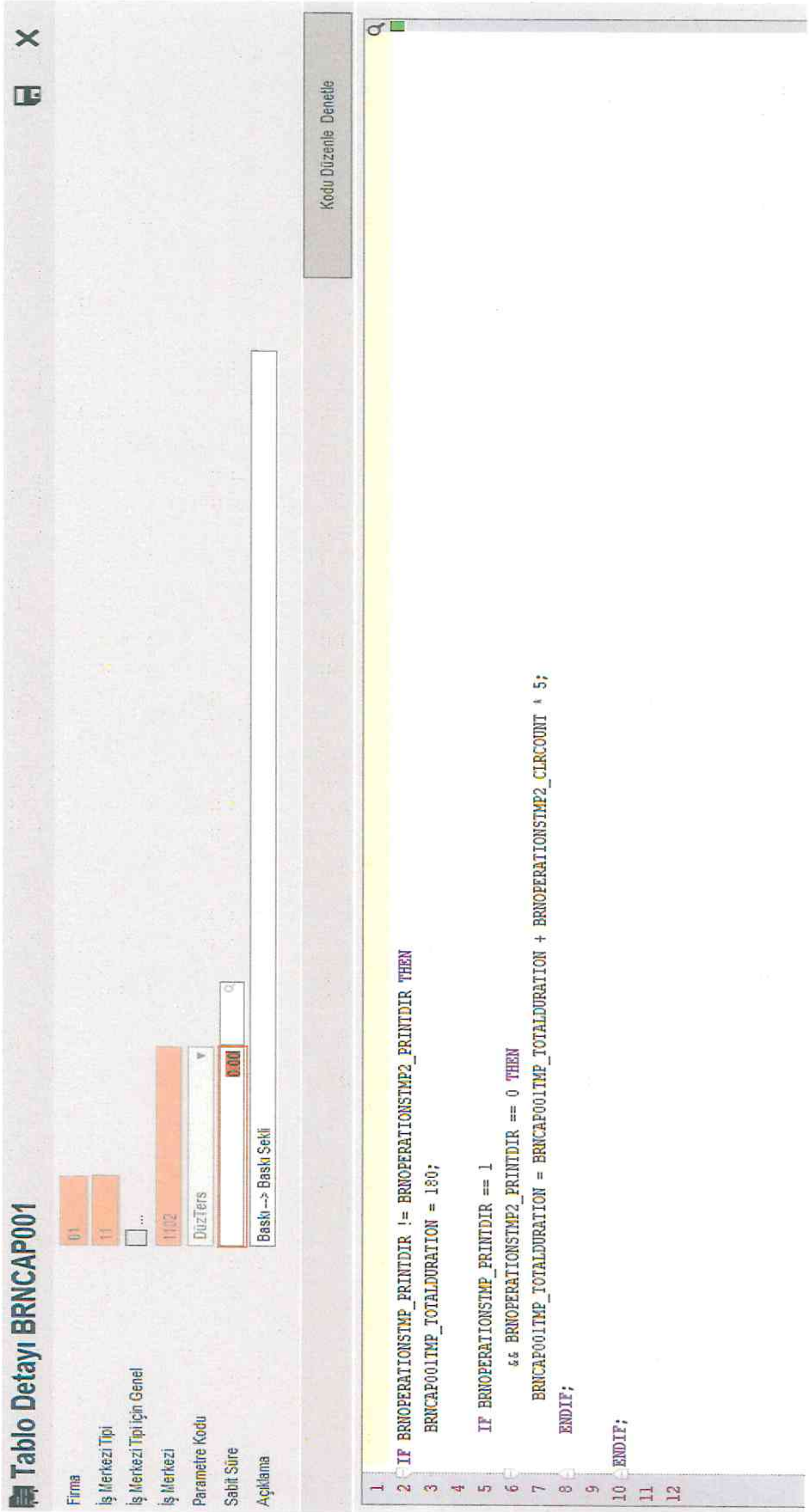


Figure 4.4. Detail Display of Setting Table

BRNOPERATIONSTMP and BRNOPERATIONSTMP2 are two virtual tables containing all the required parameters of all operations being scheduled. In this example, for 42 operations on 1109, all the information needed in the production order, order and PIF detail is kept in two tables. In these two tables, SETUPKEY fields are assigned as virtual values for 42 operations.

During the creation of the IASCAP002 data, 42 x 42 = 1764 data are placed in the BRNOPERATIONSTMP2 cycle in the BRNOPERATIONSTMP loop.

Therefore, the transition of each other is calculated using BRNCAP001 formulas.

For example; Let's calculate the transition from set up type 11090001 to type 11090002. The formula starts automatically taking 180 minutes, if the printing aspects are not the same. As a second condition, if the reverse printing is switched from flat printing (the number of engraving in the new job) x 5 minutes is added.

As a result of running the formula of 5 registered parameters, the maximum value calculated is added to the IASCAP002 table as set up time of this transition. After this data is formed, which order will be optimal is determined by the Giffler & Thompson Algorithm of the CAP module and the order of the works is done.

As an example, consider the parameters of slitting machines. They are generally entered parameters on a work centre type basis, as they are valid formulas for all slitting machines.


BRNCAP001

Firma: 01 İş Merkezi Tipi: 12

	FK	İş Merkezi Tipi	Genel?	İş Merkezi	Açıklama
1	01	12	☑	%	Dilme --> Bıçak Tipi
2	01	12	☑	%	Dilme --> Bobin Dilme Eni
3	01	12	☑	%	Dilme --> Göbek Çapı
4	01	12	☑	%	Dilme --> Malzeme
5	01	12	☑	%	Dilme --> Sarım Yönü
6	01	12	☑	%	Dilme --> Set Sayısı

Figure 4.5. Display of Setting Table for Slitting Process

Figure 4.6 Detail Display of Setting Table for Slitting Process

It can be seen that the TBLIASCAP002 virtual table was formulated by adding it to the set up time.

In the company Capacity Planning concept, custom improvements were made on CANIAS CAP module and the CAP module was worked on standard priority rules (IASCAP001). At this point, detailed improvements have been made to optimize set up times as the most critical priority rule.

IASCAP001 – Priority Rules

It is the setting table in the CAP module standard.

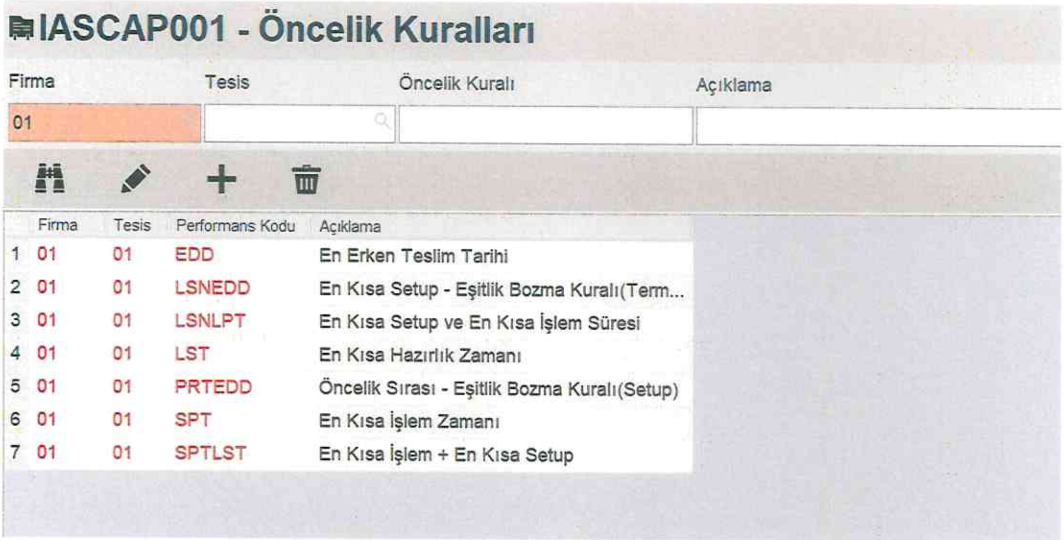


Figure 4.7. Prioritization Rules

It works on an algorithm based on minimizing PERFORMANCEVALUE and PERFORMANCEVALUE2 values in formula fields.

LSNEDD - Shortest Set up - If you look at the tie breaking example;

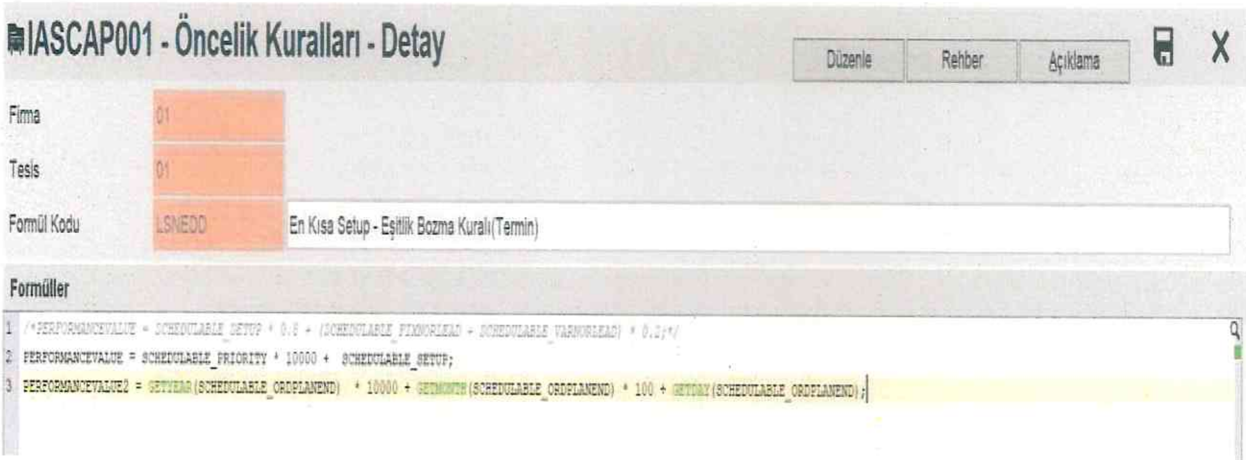


Figure 4.8. Prioritization Code Structure

The priority performance value tries to minimize the total set up time of the scheduled works (Schedulable virtual table is created in Brnscheduler class, in standard structure).

The priority performance value formula:

Performancevalue = Schedulable_Priority * 10000 + Schedulable_Setup

Priority is the priority area of the PRDT01 detail of the work orders with a value between 0-99. The reason it is directly formulated is that the works to be included in the schedule, which are stated in column Ç on the operations page developed specifically for company on CAPT01, are primarily requested to be scheduled among themselves.

The details are explained below.

The logic requested at this point; the operations listed in CAPT01 according to the work order operations table (IASPRDOPR), are included in the tasks to be scheduled under the planning user control, together with the new status fields created. It is managed via the column indicated below

Here, the most basic expectation of real life problems is customer satisfaction and customer satisfaction is ensured provided there is a balance between quality, delivery and cost. All three parameters affect the algorithm. If the product is planned in a good time period, the production line will not be hasty and the error rate will decrease. At the same time, the reduction of this error rate will affect the cost, and the most fundamental point is the timely delivery of the product to the customer.

At this point, we run the G&T algorithm under the EDD (Earlies Due Date) harness, taking into account the customer delivery time, and then LST (Least Set Up Time) to optimize the long set up times of the bag machines. At the last stage, we run the LSNEDD (Least Set up and Earliest Due Date) harness to ensure the algorithm is balanced. You can see the results of these three harnesses from the graph below.

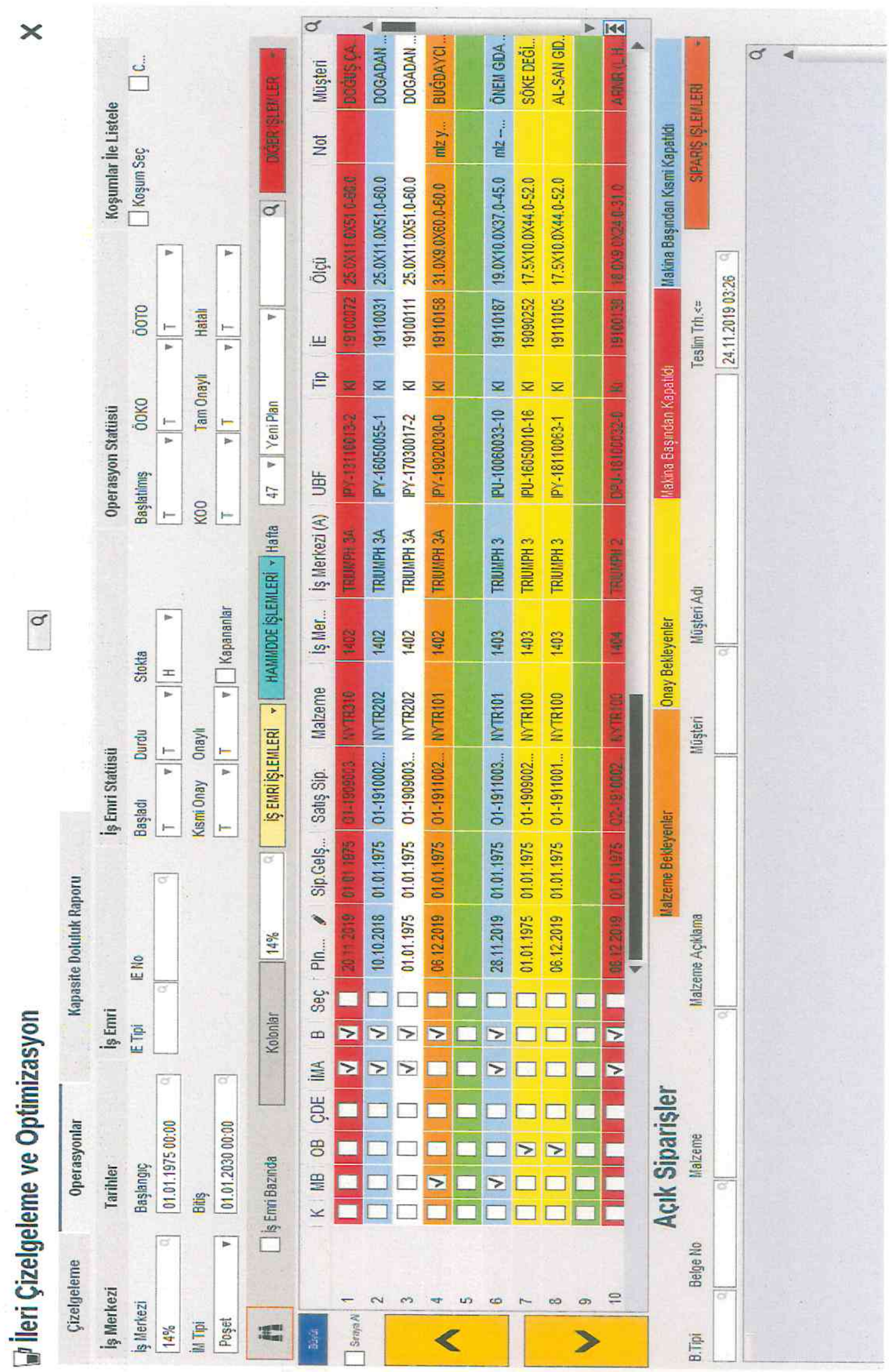


Figure 4.9. Selection of Jobs to be Included in Scheduling

The planning user will include the desired jobs in the scheduling, the included jobs will be scheduled among themselves and according to priority dates, and the remaining jobs in the list will be scheduled after them so that the set up times are minimized. Once the jobs to be scheduled are received, in the BRNSCHEDULER class Ç column, virtual values are assigned to marked operations (IASPRDOPR_ISINCSCHEDULE) with priority values 1 and unmarked jobs with priority values 99. Therefore, in this example, PRIORITY is directly contained in the PERFORMANCEVALUE value. Thus, the existing scheduling algorithm can be used without changing.

The second priority performance value formula;

Performancevalue2 = Getyear(Schedulable_Ordplanend) * 10000 + Getmonth(Schedulable_Ordplanend) * 100 + Getday(Schedulable_Ordplanend)

The second priority performance value is defined so that if the system attempts to minimize the set up value during the ranking of jobs, if the set up encounters 2 alternatives with the same set up values, the deadline date will take priority. The specified deadline (ORDPLANEND) is the planned end date in the production order detail. When the production order is opened, it is assigned to this field as the planned end date of the day before the deadline entered in the order.

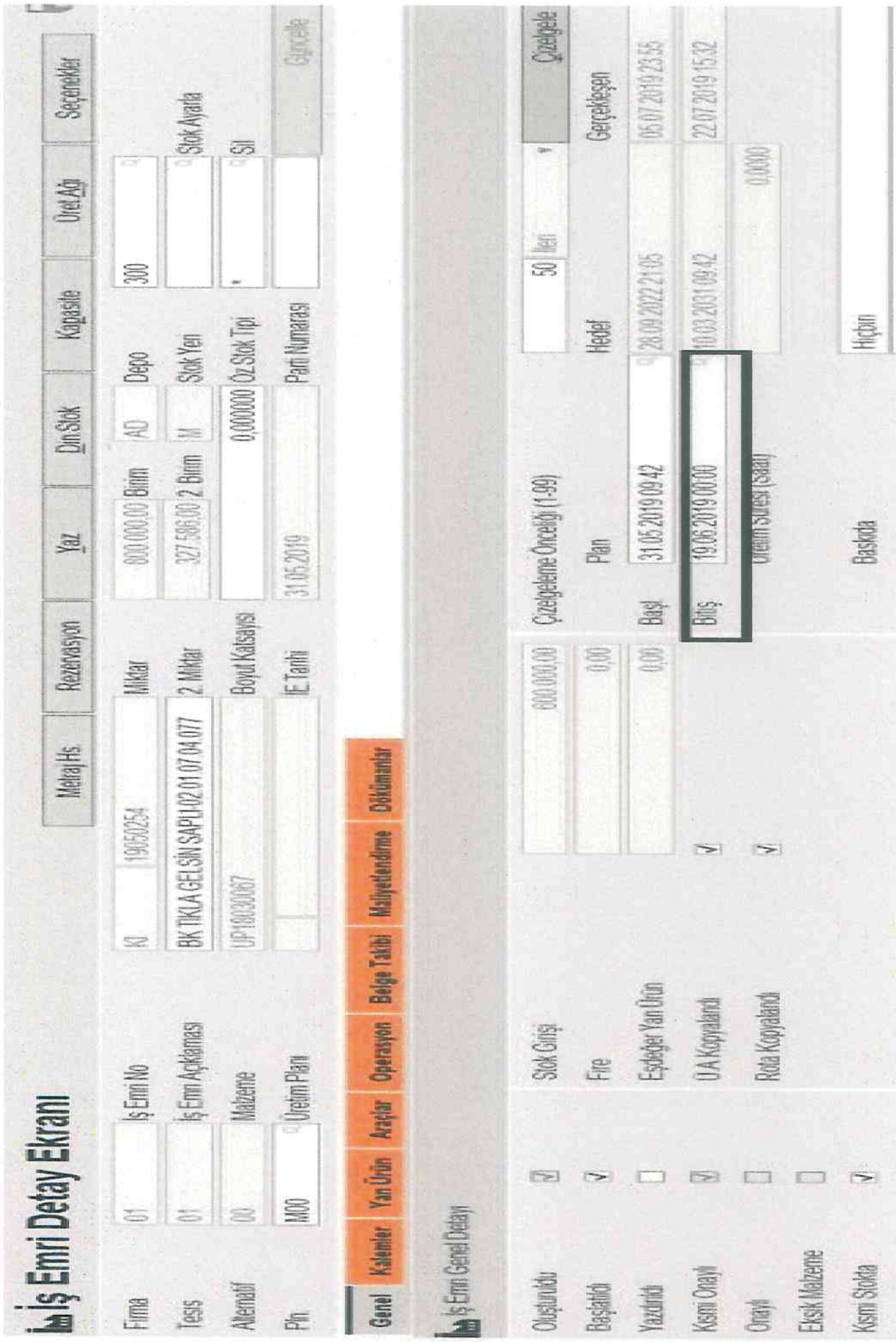


Figure 4.10 Updating Production Order Data

The criteria according to which operations will be sorted are determined by this priority rule selected during scheduling in CAPT01. As a note for example; Since customer information is included in production orders, the jobs of the related customers can be moved forward or backward in scheduling with the formula to be given in IASCAP001 detail. However, LSNEDD, which has already been formulated according to the available information, seems to be the most suitable alternative for company in the analyses.

LST – With the example of the Shortest Set up Time, set up optimization can be performed with a single performance value.

Performancevalue = Schedulable_Setup

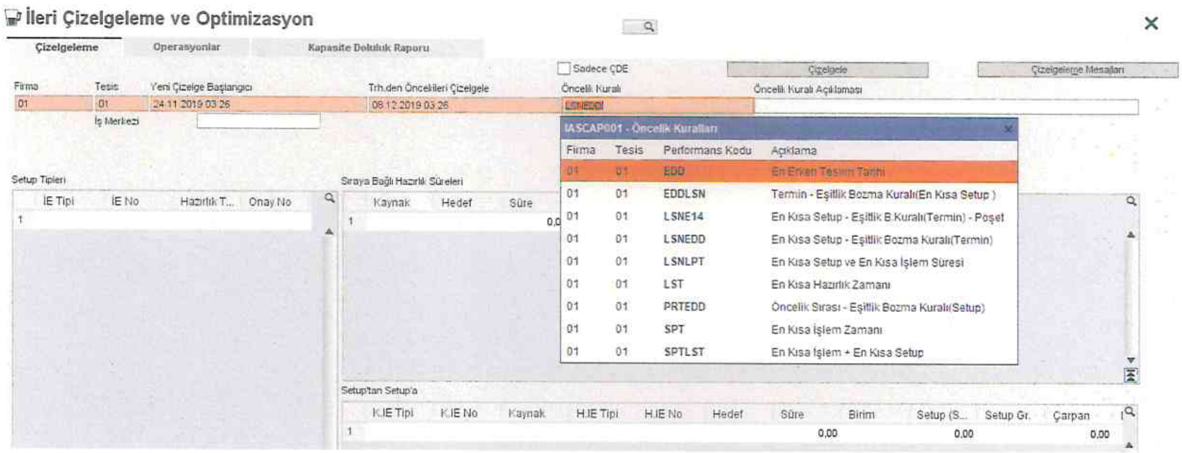


Figure 4.11 Shortest Set Up Time Algorithm

The most crucial issue in scheduling for company is to minimize set up times. The scheduling priority rule to be selected must contain the LST concept in any case.

LST – The shortest set up time (Schedulable Setup) performance value requires the use of the IASCAP002 - Sequential set up times setting table.

In CAPT01 temporarily, the following example has been run by selecting LSNEDD for 1109 so that the work can be performed on a work centre basis until the work is verified and completed.

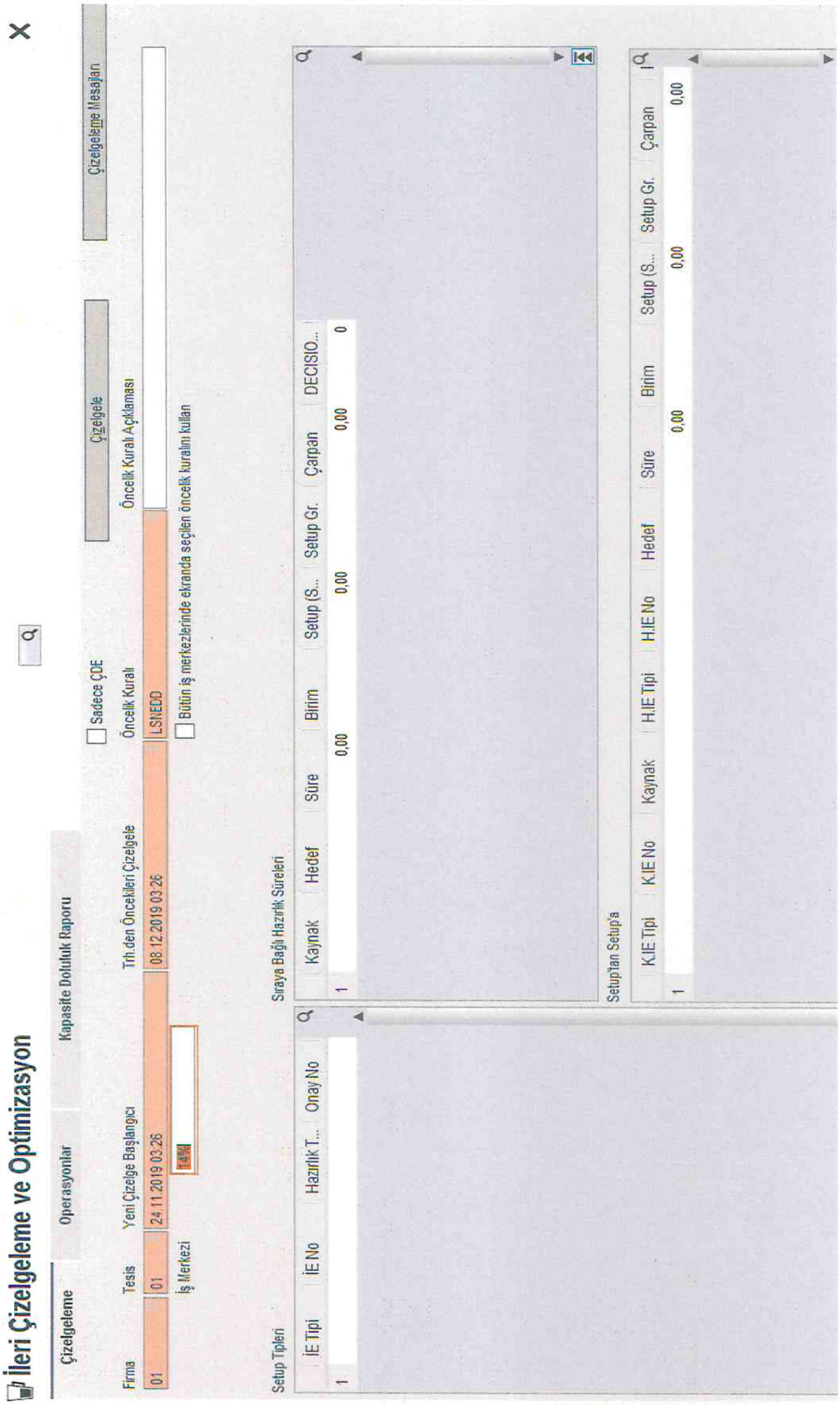


Figure 4.12 Running System with Shortest Set Up Time

The time and date information of the operations are not updated directly and records are kept in a separate table in the database for the relevant condition. Although it is kept in the database, the current operations are constantly changing as the CAP process is constantly updated with the progress of production. Therefore, the only possible way to list related condition is to kept the operating screen is on. As a result of the process, the virtual IASROU004 and IASCAP002 data for that condition can be seen on the screen.

In set up types, by double clicking on the line, the transitions of that production order to other production orders can be controlled by placing them directly to the right.

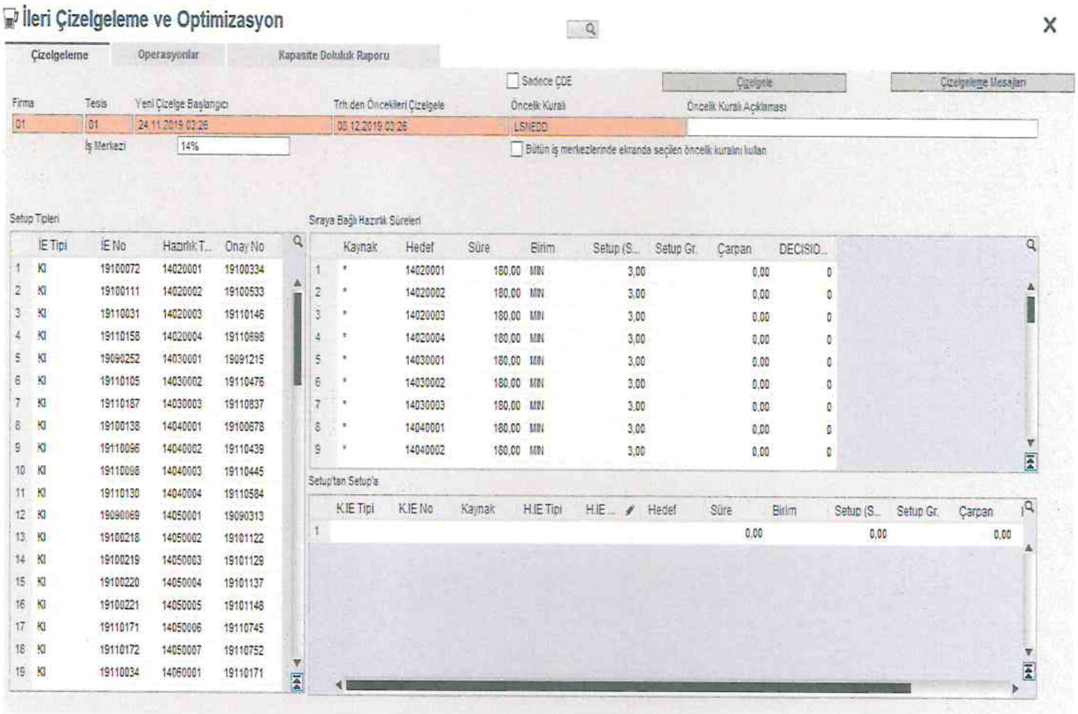
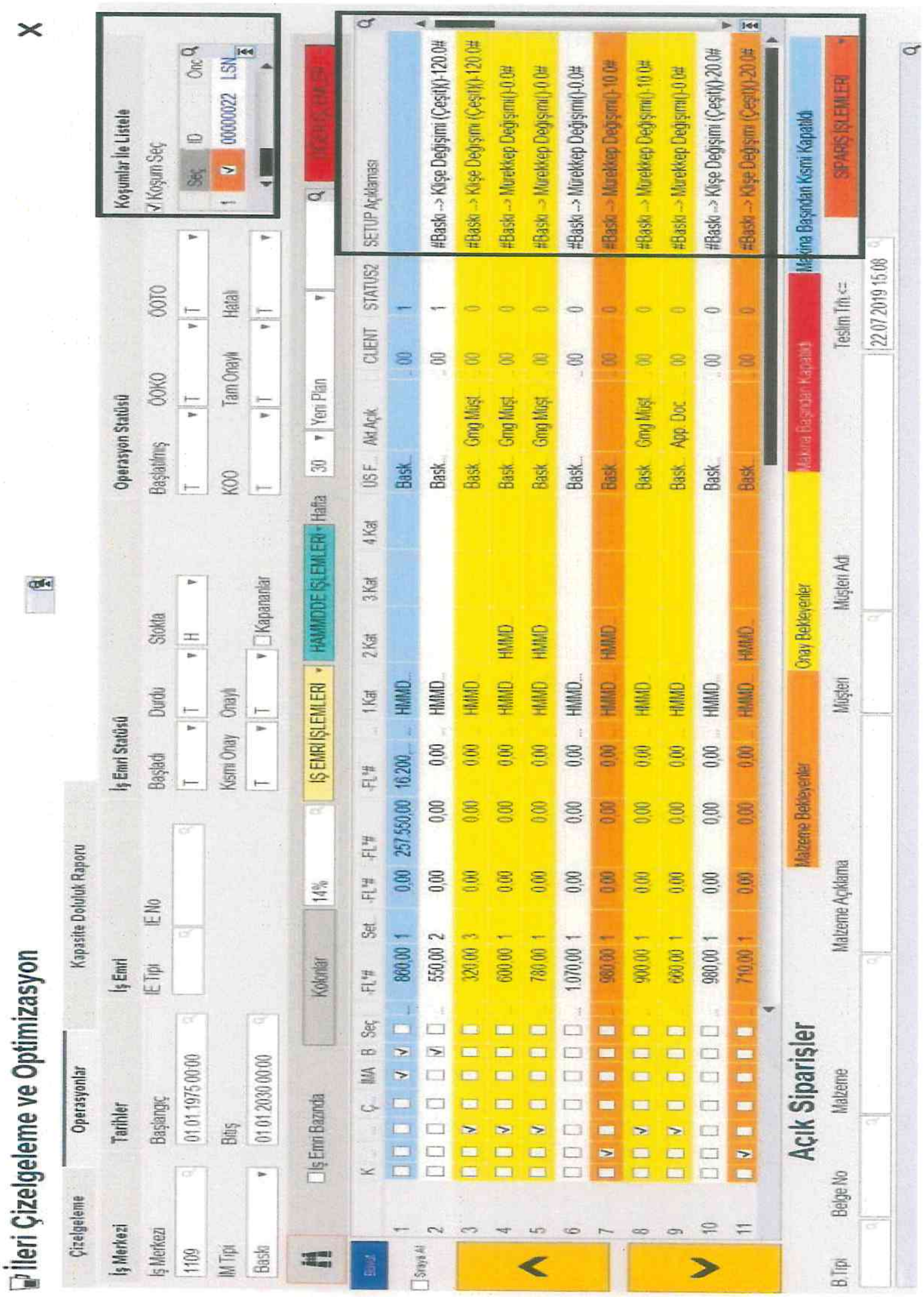


Figure 4.13 Display of Set Up Times by Production Order

In the Operations tab, the statuses can be selected and the results can be listed. For set up optimization, the column can be seen at the end, which explains why the next job is selected in the ranking. At the last point, if the selected condition is the desired result, if the rows after the listing are saved, the order proposed by the system is recorded. This can be achieved after system verification and necessary corrections are made.



*F***igure 4.14**. Display of Scheduled Operations

4.2 Results and Analysis

The Scheduling results were tested through the Canias ERP system and these test results were evaluated through three algorithms. As a strategy, especially EDD (Earlies Due Date) can increase customer satisfaction while increasing internal production costs. However, while LST (Least Set Up Time) tries to minimize costs, it can also create problems in terms of customer orientation. The results of these two algorithms will be compared with the results of LSNEDD (Least Set Up and Earliest Due Date) algorithm which can balance both customer orientation and costs. As a result, the algorithm to be used within the company will be revealed.

4.2.1 Comparison Scheduling Rules

Eighty jobs for analysis were selected for thirteen bag lines for a 10 months. The characteristics of each bag line vary according to the type of work. In bag machines, set up times are long and the number of set up types varies according to the measurements. The most important criterion here is the ratio of the bag. When the size changes, the bag machine is set up

Here, the most basic expectation of real life problems is customer satisfaction and customer satisfaction is ensured provided there is a balance between quality, delivery and cost. All three parameters affect the algorithm. If the product is planned in a good time period, the production line will not be hasty and the error rate will decrease. At the same time, the reduction of this error rate will affect the cost, and the most fundamental point is the timely delivery of the product to the customer.

At this point, we run the G&T algorithm under the EDD (Earlies Due Date) harness, taking into account the customer delivery time, and then LST (Least Set Up Time) to optimize the long set up times of the bag machines. At the last stage, we run the LSNEDD (Least Set up and Earliest Due Date) harness to ensure the algorithm is balanced. You can see the results of these three harnesses from the graph below.

Table 4.4 Comparison of Scheduling Rules

	Line	Production Order (Pcs)	Total Set Up Time (Hour)	Total Production Time (Hour)	C_{max} (Hour)	Tardiness (Hour)
LST	TR-3A	13	41.28	264	305.3	44
EDD			53.17		317.2	7
LSNEDD			49.88		313.9	9
MANUEL			57.58		321.6	65
LST	TR-5QT1	12	38.25	278	316.3	74
EDD			50.36		328.4	4
LSNEDD			45.35		323.4	7
MANUEL			52.24		330.2	79
LST	TR-5QT2	15	45.89	298	343.9	58
EDD			63.45		361.5	6
LSNEDD			48.65		346.7	8
MANUEL			67.39		365.4	62
LST	TR-3MR	13	37.41	326	363.4	62
EDD			55.98		382.0	6
LSNEDD			41.48		367.5	8
MANUEL			58.96		385.0	68

The factory's manual scheduling using manual heuristics was carried out using G&T algorithm with more iterations. Thus, some algorithms used in the standard were compared with the planning department. However, the LSNEDD rule was used, taking into account the priority of the G&T algorithm, set up, delivery time and the priority of the two algorithms. In these algorithms, the change of $C_{max}$was examined on line basis. The LST algorithm comes to the forefront in lines with set up time, especially when the delivery date approaches, regardless of the set up cost dramatically changes C_{max}.

LST is the give best results for C_{max} but it is also increase the tardiness so it give us unsatisfied customer. EDD rule is about the customer satisfaction but it is also increase the set up time and it is also increase the cost of product. The LSNEDD rule offers a balanced strategy between customer satisfaction and set up time. Therefore, this algorithm seems to be very convenient for this factory.

Table 4.5 Comparison of Scheduling Rules Total C_{max} and Tardiness

	C_{max} (Hour)	Tardiness(Hour)
LST	1328.8	238
EDD	1389.0	23
LSNEDD	1351.4	30
MANUEL	1402.2	274

On the other hand, the manual algorithm is relatively weak. As you can see in the graph above, it has produced even worse results than the worst-performing algorithm. Comparing LSNEDD with manual scheduling methods, a C_{max} improvement difference of 3% and a tardiness of 12 times better results.

It is seen that such scheduling systems remain inappropriate with the use of today's technology. The dynamic structure of the production environment is constantly changing the schedule table. When these instant reflexes are not answered, scheduling algorithms cannot be used effectively in production environments. For example, TR-3A, even if the raw material of product is not suitable, if the work proceeds slower than expected, the table should update itself automatically with this data and it should be planned with the expected stops not only due to similar failures caused by the machine in some similar works. The company uses ERP and MES systems together and has successfully completed horizontal integration. Instant changes are informed visually by message and mail as below. At the same time, these two systems transfer instant data by talking to each other. The size of monthly data collected from MES system is 20 GB.

The data is analysed using R 3.6.1 software and clustering technique. The related software is free software and some libraries have been used, but generally the algorithm design process belongs to us. This software is used as an analysis tool between MES and ERP. This software makes the data it receives from MES database meaningful and then transfers it to ERP database.

4.2.2 Data Mining and Updating System

Cluster analysis with R is performed in two steps. The first step is to prepare the data. This means that the data is import from the computer to the R software. The second stage is to cluster the data according to the specified criteria.

1. Data Preparation Code

```
# Prepare Data
mydata <- na.omit(mydata) # listwise deletion of missing
mydata <- scale(mydata) # standardize variables
```

2. Partitioning

```
#Determine number of clusters
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum (kmeans(mydata, centres=i)$withinss)
plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within
groups sum of squares")

#K-Means Cluster Analysis
fit<-kmeans(mydata,5) #5 cluster solution
#get cluster means
aggregate(mydata,by=list(fit$cluster),FUN=mean)
#append cluster assignment
mydata <- data.frame(mydata, fit$cluster)
```

The most important parameter in cluster analysis is how many clusters will be processed. Here, the most important criterion for the production of bags is the length of the bags. When this length changes, bag machines are adjusted. There are eight different width types used in the above analysis. We analyse the data according to these eight types. The criterion I use the cluster algorithm is the length which is decisive for production orders. The parameters of the work orders are grouped according to these dimensions. According to this grouping, set up times, cycle times of production orders, machine failures and repair times are updated. Set up times are updated in the light of the data from R software in Figure 4.7. Over 10% improvement in set up times after updating from these areas. A 5% reduction was observed in cycle times

Table 4.6 Comparison of Scheduling Rules After Updating Parameters

	Line	Production Order (Pcs)	Total Set Up Time (Hour)	Total Production Time (Hour)	C_{max} (Hour)	Tardiness (Hour)
LST	TR-3A	13	35.622	264	284.6	29.1
EDD			46.242		294.7	0.0
LSNEDD			41.697		290.4	0.0
MANUEL			47.295		295.7	39.2
LST	TR-5QT1	12	33.795	278	296.2	52.7
EDD			43.398		305.3	0.0
LSNEDD			38.835		301.0	0.0
MANUEL			45.135		307.0	54.8
LST	TR-5QT2	15	40.68	298	321.7	33.5
EDD			55.8		336.1	0.0
LSNEDD			43.2		324.1	0.0
MANUEL			58.5		338.7	35.7
LST	TR-3MR	13	31.5	326	339.6	38.6
EDD			49.68		356.9	0.0
LSNEDD			35.1		343.0	0.0
MANUEL			50.589		357.8	41.5

In Table 4.6 total production time and C_{max} and Tardiness decrease. LSNEDD and EDD algorithm tardiness is zero now. But EDD is high set up time and it means extra cost for production so LSNEDD is give us better result after update the system.

Table 4.7 Comparison of Scheduling Rules Total C_{max} and Tardiness After Updating Parameters

	C_{max} (Hour)	Tardiness(Hour)
LST	1242.2	153.8872
EDD	1293.1	0
LSNEDD	1258.6	0
MANUEL	1299.1	171.2331

As you see Table 4.7 total C_{max} is less than before updating system. After set up time and cycle time update the C_{max} is decrease % 6.3 for LSNEDD algorithm and total tardiness is zero for this algorithm. If generally examine Table 4.7, the C_{max} and tardiness decreasing. In these production system the best rule is LSNEDD.

CHAPTER 5
CONCLUSION AND FUTURE WORK

The development of real-time scheduling algorithms will enable us to provide the flexibility to respond to instantaneous changes in production environments. Adaptive scheduling algorithms contribute to the achievement of the company's key performance index for production department. Real time scheduling allows us to constantly control production parameters and make instant decisions.

Moving the scheduling process to the ERP system and eliminating all manual processes within the company caused two engineers to fall to one engineer. An engineer who constantly updates the data in Ms excel is now evaluated in another department along with the system self-data update. In addition, instant data update support is provided to support the scheduling algorithm integrated into the system. Thus, the data that is critical for scheduling will be kept up to date with this structure on a regular basis. This will be provided by MES with the data of the structure.

This helps us to use raw material and labour resources in production more efficiently. Optimal use of labour and raw materials will reduce production costs and contribute to the company's quality production and R & D processes. Companies that produce on time, high quality and cost effective products will step forward in the current conditions. Adaptive scheduling algorithms are related to these three important parameters. These algorithms, which affect these three parameters and enable us to act flexibly in the production environment, make a significant contribution to the logistics processes. The algorithm written in this thesis points to a new phase of integrated production systems and logistic process. At the beginning, we only schedule jobs and try to comply with the deadlines of our schedule, and now we are able to analyse production environment and make instant decisions with production management systems. These instant decisions directly affect the macro and micro decisions of the company. Such production systems can respond to market demands more quickly. At the same time, the rapid development of technology and the development of self-decision systems show us that these systems will further develop. Considering that all processes from order to product delivery will be digitized, the integration of such start up systems is very important for the systems to be integrated in the future. The system used in this thesis is to generate meaningful data from big data by using periodical updating of the parameters used by the existing scheduling algorithm and data mining methods. With the development of systems, the size and diversity of the data collected has increased. In this increasing data, it is becoming increasingly important to extract the data in the meaning we want and use it for production. One of the important points in the thesis is the effect of data generated by digitized processes on scheduling production processes.

These data are the data of integrated production systems that are digitalized and communicating continuously and meaningful data are extracted from these data that will affect production or production planning. With this thesis, further studies will be carried out on improving production processes, production planning and digitalization. These studies will explore the system behaviour in high capacity utilization in order to reveal greater potential for flexibility and responsiveness in different production conditions.

REFERENCES

1. Adams, J., E. Balas, and D. Zawack, *The Shifting Bottleneck Procedure for Job Shop Scheduling.* Management Science, 1988. **34**(3): p. 391-401.
2. Giffler, B. and G.L. Thompson, *Algorithms for Solving Production-Scheduling Problems.* Operations Research, 1960. **8**(4): p. 487-503.
3. Denkena, B., M.-A. Dittrich, and S. Wilmsmeier, *Automated production data feedback for adaptive work planning and production control.* Procedia Manufacturing, 2019. **28**: p. 18-23.
4. Wang, C.L., et al., *Mining scheduling knowledge for job shop scheduling problem.* IFAC-PapersOnLine, 2015. **48**(3): p. 800-805.
5. Denkena, B., J. Schmidt, and M. Krüger, *Data Mining Approach for Knowledge-based Process Planning.* Procedia Technology, 2014. **15**: p. 406-415.
6. Denkena, B. and T. Mörke, *Cyber-Physical and Gentelligent Systems in Manufacturing and Life Cycle*. 2017.
7. Denkena, B., L.-E. Lorenzen, and J. Schmidt, *Adaptive process planning.* Production Engineering, 2012. **6**(1): p. 55-67.
8. Govindaraju, R. and K. Putra, *A methodology for Manufacturing Execution Systems (MES) implementation*. Vol. 114. 2016. 012094.
9. Frank, A.G., L.S. Dalenogare, and N.F. Ayala, *Industry 4.0 technologies: Implementation patterns in manufacturing companies.* International Journal of Production Economics, 2019. **210**: p. 15-26.
10. Türkeş, M.C., et al., *Drivers and barriers in using industry 4.0: A perspective of SMEs in Romania.* Processes, 2019. 7(3).
11. Broy, M., M. Cengarle, and E. Geisberger, *Cyber-Physical Systems: Imminent Challenges*. Vol. 7539. 2012. 1-28.
12. Ciurana, J., et al., *A model for integrating process planning and production planning and control in machining processes.* Robotics and Computer-Integrated Manufacturing, 2008. **24**(4): p. 532-544.
13. M. Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, *From Data Mining to Knowledge Discovery in Databases*. Vol. 17. 1996. 37-54.
14. Mohan, J., K. Lanka, and A.N. Rao, *A Review of Dynamic Job Shop Scheduling Techniques.* Procedia Manufacturing, 2019. **30**: p. 34-39.
15. Shabtay, D., *Scheduling and due date assignment to minimize earliness, tardiness, holding, due date assignment and batch delivery costs.* International Journal of Production Economics, 2010. **123**(1): p. 235-242.
16. Brucker, P. and S. Knust, *Complex Scheduling (GOR-Publications)*. 2006: Springer-Verlag.
17. Pinedo, M.L., *Scheduling: Theory, Algorithms, and Systems*. 2008: Springer Publishing Company, Incorporated. 678.
18. Baker, K. and D. Trietsch, *Principles of Sequencing and Scheduling*. 2018.
19. Pinedo, M. and X.A. CHAO, *Operations scheduling with applications in manufacturing and services*. 1999: McGraw-Hill Companies.
20. Birogul, S. and C. Elmas, *Genetik Algoritma Yaklaşımıyla Atölye Çizelgeleme*. 2005.
21. Brucker, P., *Scheduling Algorithms*. Vol. 47. 2004.

22. Cheng, M.G.R., *Genetic Algorithms and Engineering Design.* WİIey, 1997.
23. R. Walters, J., *Heuristic Scheduling Svsteins. By THOMAS E. MORTON and DAVID W. PENTICO (Wiley, 1993).* International Journal of Production Research - INT J PROD RES, 1995. **33**: p. 1787-1788.
24. Ponnambalam, S.G., N. Jawahar, and B.S. Girish, *Giffler and Thompson Procedure Based Genetic Algorithms for Scheduling Job Shops*, in *Computational Intelligence in Flow Shop and Job Shop Scheduling*, U.K. Chakraborty, Editor. 2009, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 229-259.
25. Perzyk, M., et al., *Comparison of data mining tools for significance analysis of process parameters in applications to process fault diagnosis.* Information Sciences, 2014. **259**: p. 380-392.
26. Laney, *3-D Data Management: Controlling Data Volume, Velocity and Variety.* 2001
27. Reuter, C. and F. Brambring, *Improving Data Consistency in Production Control.* Procedia CIRP, 2016. **41**: p. 51-56.
28. Cheng, Y., et al., *Data and knowledge mining with big data towards smart production.* Journal of Industrial Information Integration, 2018. **9**: p. 1-13.
29. Köksal, G., İ. Batmaz, and M.C. Testik, *A review of data mining applications for quality improvement in manufacturing industry.* Expert Systems with Applications, 2011. **38**(10): p. 13448-13467.
30. Piatetsky-Shapiro, G. *Data mining and knowledge discovery in business databases.* in *Foundations of Intelligent Systems.* 1996. Berlin, Heidelberg: Springer Berlin Heidelberg.
31. Begoli, E. and Edmon, *Design Principles for Effective Knowledge Discovery from Big Data.* 2012.
32. Demetrescu, M. and S. Hacıoğlu Hoke, *Predictive regressions under asymmetric loss: Factor augmentation and model selection.* International Journal of Forecasting, 2019. **35**(1): p. 80-99.
33. Sagiroglu, S. and D. Sinanc. *Big data: A review.* in *2013 International Conference on Collaboration Technologies and Systems (CTS).* 2013.
34. Mason, R.J., M.M. Rahman, and T.M.M. Maw, *Analysis of the manufacturing signature using data mining.* Precision Engineering, 2017. **47**: p. 292-302.
35. Della Croce, F., R. Tadei, and G. Volta, *A genetic algorithm for the job shop problem.* Computers & Operations Research, 1995. **22**(1): p. 15-24.
36. Yamada, T. and R. Nakano, *A Genetic Algorithm Applicable to Large-Scale Job-Shop Problems.* Vol. 2. 1992. 283-292.
37. Cavalcante, I.M., et al., *A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing.* International Journal of Information Management, 2019. **49**: p. 86-97.
38. Boone, T., et al., *Perspectives on supply chain forecasting.* International Journal of Forecasting, 2019. **35**(1): p. 121-127.
39. Vazquez, M. and L.D. Whitley. *A comparison of genetic algorithms for the dynamic job shop scheduling problem.* in *Proceedings of the 2nd Annual Conference on Genetic and Evolutionary Computation.* 2000. Morgan Kaufmann Publishers Inc.
40. Anderberg, M.R., *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks.* Vol. 19. 2014: Academic press.
41. Lemey, P., M. Salemi, and A.-M. Vandamme, *The phylogenetic handbook: a*

practical approach to phylogenetic analysis and hypothesis testing. 2009: Cambridge University Press.