

1. Introduction

TrackAnalytics is a company that offers data analysis to its customers. Sometimes, it uses customer's data (with no cost) to undertake all analysis, but many times it gets data from third party sources (with a cost) in order to perform its work. The data mining process ends with a report having suggestions to address the customer's needs. For example, if Ford wants to know if the Fusion model needs a new design, TrackAnalytics can buy data from Facebook and Twitter to identify any users' posts mentioning that other companies are better for having great designs, or that Ford is not renewing its designs very often as a bad thing. To do this kind of work, the company resorts to data scientists to analyze all data and issue a final report, which is delivered to the client company as a product.

2. Purpose

This project intends to develop a web scraper application (software that can collect data from web pages) to offer an alternative way to collect data from the internet as a means to minimize data scientists' dependency on third party APIs (source of online data controlled by others) made available online. As a result, data scientists will have access to their own data which can guarantee source of data to undertake a desired analytical study without paying or expecting availability from third parties. Once successful, this project will provide autonomy and cost reduction on future data collection and study.

3. Objectives

In an team effort, the development of a server and a client side web application will deliver a simple user friendly tool for collecting data from the web that need to be usable by anyone with data science skills within TrackAnalytics.

4. Overview

The new platform project will build a web scraper web application to minimize dependency on data collection from third party source, making future projects more autonomous and less costly given that any data scientist will be able to apply the tool to gather data from the internet in order to undertake a required analytical study.

The projection is that this project will take around a month comprising two milestones. The first one will take the first half of the schedule having the development of the server and client sides in parallel, and the second one will take the remaining allocated time involving the integration between server and client sides. The main constraint to this project are personnel skills and time, and the approach to mitigate these aspects are the implementation of in-house data scientists' domain knowledge technology and project management tools in order to deliver the expected outcome on schedule, on scope, on budget, and fulfilling the requirements. Because the project is internal, there are some flexibilities like the possibility to compromise on the beauty of the user interface, but the deliverable date is June 26th 2017, for any new study starting after this date will be using the delivered tool. Given the small size of the company, it is assumed that people involved on the project will be working 25% of their regular time on the project and the other 75% on customer's work. The data scientists work 8 hours per weekday. Any need to keep the project on schedule or to expedite it will happen through overtime work, which incurs on a increase of 50% on the payment rate.

5. Project Activity Analysis

The first activity will be a meeting where the data scientists will define the technologies and design the project will have. The meeting will be lead by the senior data scientist attending it. The meeting will take 7 days.

Having in mind a big picture that the project has a server side and a client side, the development of both sides will start in parallel by the data scientists after the meeting. Half of the data scientists on the server side and another half on the client side. A WBS will highlight the people who will be involved on

each activity. Basically, the activities will have all data scientists, data scientists on the server team, and data scientists on the client team. Among these three groups, the senior data scientist of each group will lead the activity. The PM(Project Manager) must be informed on the developments by the senior of each group who will be designated "lead".

The server side development will have 4 activities happening concurrently. One will be to code the configuration of the software. This activity will take 7 days. Another will be to code the models which prepares the database connection and use. This will take 7 days. Another will be to code the controllers which transition the data. This will take 7 days. Another will be to code the routes which is the data link. This will take 7 days.

The client side development will have 4 activities happening concurrently. One will be to code the view which is the web page. This will take 7 days. Another will be to code the view's controllers which transition the data. This will take 7 days. Another will be to code the view's services which process the data. This will take 7 days. Another will be to code the style sheet which format the web page. This will take 7 days.

Once the server and client sides are completed, all developed code will be transferred to a Git repository by the data scientists in order to be widely available for integration. This activity will take 2 days. After this, the data scientists who developed the client side code will work on the client side integration which make each piece of code to talk to each other. This activity will take 7 days. In parallel, the data scientists who developed the server side code will work on the server side integration which make each piece of code to talk to each other. This activity will take 7 days.

Once both integration activities are completed, the data scientists who integrated the client side will test the server side for quality assurance. This will take 2 days. In parallel, the data scientists who integrated the server side will test the client side for quality assurance. This will take 2 days.

After both testings, the data scientists who integrated the server side will integrate the server and client sides which make both sides to talk to each other. This will take 7 days. After this, the data scientists who integrated the client side will test the final software for quality assurance. This will take 2 days.

The time estimates followed a bottom up approach where three times will be considered such as an optimistic, a most likely, and a pessimistic. To obtain these three times, the data scientists based their calculations on previous duration of coding work developed within the company. The optimistic time corresponds to a task duration "a" or lower that happened less than 1 percent of the time. The pessimistic time corresponds to the same task duration "b" or greater that happened less than 1 percent of the time. The most likely time corresponds to the same task duration "m" referent to the mode of the distribution. In this company, the optimistic time is very related to periods when the company had less volume of client's work while periods presenting a high volume of client's work is very related to the pessimistic time. The table presenting the estimates is as follows:

Activity ID	Optimistic Time (a) (days)	Most Probable Time (m) (days)	Pessimistic Time (b) (days)	Expected Time $E(t)=(a+4*m+b)/6$ (weeks)	Variance $\sigma^2=[(b-a)/6]^2$
A	3	7	9	6.6	1
B	3	7	9	6.6	1
C	3	7	9	6.6	1
D	3	7	9	6.6	1
E	3	7	9	6.6	1
F	3	7	9	6.6	1
G	3	7	9	6.6	1

Activity ID	Optimistic Time (a) (days)	Most Probable Time (m) (days)	Pessimistic Time (b) (days)	Expected Time $E(t)=(a+4*m+b)/6$ (weeks)	Variance $\sigma^2=[(b-a)/6]^2$
H	3	7	9	6.6	1
I	3	7	9	6.6	1
J	1	2	3	2	0.111
K	3	7	9	6.6	1
L	3	7	9	6.6	1
M	2	2	2	2	0
N	2	2	2	2	0
O	3	7	9	6.6	1
P	2	2	2	2	0

The Work Breakdown Structure (WBS) was obtained following a hierarchical planning process where the data scientists resorting to sticky-notes defined the tasks needed to accomplish the project. The WBS is as follows:

WBS (Work Breakdown Structure)				
Activity ID	Activity Name	Duration	Predecessor or	Involved People
A	Define the project's technologies and design	7 days	None	Data Scientists
B	Code the server side configuration	7 days	A	Data Scientists "server"
C	Code the server side models	7 days	A	Data Scientists "server"
D	Code the server side controllers	7 days	A	Data Scientists "server"
E	Code the server side routes	7 days	A	Data Scientists "server"
F	Code the client side view	7 days	A	Data Scientists "client"
G	Code the client side view's controllers	7 days	A	Data Scientists "client"
H	Code the client side view's services	7 days	A	Data Scientists "client"
I	Code the client side view's style sheet	7 days	A	Data Scientists "client"
J	Move the client and server sides to Git	2 days	B,C,D,E,F, G,H, and I	Data Scientists
K	Integrate the client side	7 days	J	Data Scientists "client"
L	Integrate the server side	7 days	J	Data Scientists "server"

WBS (Work Breakdown Structure)				
M	Test the integrated client side	2 days	K and L	Data Scientists “server”
N	Test the integrated server side	2 days	K and L	Data Scientists “client”
O	Integrate the client and server sides	7 days	M and N	Data Scientists “server”
P	Test the final software	2 days	O	Data Scientists “client”

Based on the WBS, a RACI matrix describes the level of responsibility on the project. It is as follows:

RACI Matrix					
Activity ID	PM	Meeting Lead	Server Lead	Client Lead	Data Scientists
A,J	I,C	A			R
B,C,D,E,L,M,O	I	C	A		R
F,G,H,I,K,N,P	I	C		A	R

Legend: A = Accountable, C = Consult, R = Responsible, I = Inform

The cost estimate was established applying a bottom up approach. On the resources to accomplish the project, the data scientists will be using their regular computers. As mentioned on the overview, they will be allocating 25% of their regular time to the project, and any need to put an extra effort to keep the project on schedule or to accelerate it will happen through overtime work.

The table below presents the payment average rate applied by TrackAnalytics based on the Occupational Employment Statistics (OES) data on the position known as Scientific Research and Development Services.

Resource Cost Per Unit For Coding The Software					
ID	Resource Name	Max. Units	Std. Rate	Ovt.Rate	Accrue At
1	Data Scientist	1	\$61.41/hr	\$92.11/hr	Prorated

The budget by resource follows below:

Budget By Resource For Coding The Software							
ID	Task Name	Number of Data Scientists “D”	Hour Per Day Per Person (8*.25) “H”	Resource Work Hours (D*H)	Duration (Days) “A”	Std. Rate “S”	Cost (D*H*A*S)

Budget By Resource For Coding The Software							
A	Define the project's technologies and design	8	2	16	7	61.41	6877.92
B	Code the server side configuration	1	2	2	7	61.41	859.74
C	Code the server side models	1	2	2	7	61.41	859.74
D	Code the server side controllers	1	2	2	7	61.41	859.74
E	Code the server side routes	1	2	2	7	61.41	859.74
F	Code the client side view	1	2	2	7	61.41	859.74
G	Code the client side view's controllers	1	2	2	7	61.41	859.74
H	Code the client side view's services	1	2	2	7	61.41	859.74
I	Code the client side view's style sheet	1	2	2	7	61.41	859.74
J	Move the client and server sides to Git	8	2	16	2	61.41	1965.12
K	Integrate the client side	4	2	8	7	61.41	3438.96
L	Integrate the server side	4	2	8	7	61.41	3438.96
M	Test the integrated client side	4	2	8	2	61.41	982.56
N	Test the integrated server side	4	2	8	2	61.41	982.56
O	Integrate the client and server sides	4	2	8	7	61.41	3438.96
P	Test the final software	4	2	8	2	61.41	982.56
	TOTAL						28985.52
	TOTAL + 25% Contingency						36231.9

Based on the above table, it is possible to build a table contrasting the normal cost and normal time with the crash time and crash cost. The crash time and crash cost are obtained based on the premise that employees will need to do overtime work. As stated on the Resource table, the overtime rate is \$92.11/hr. The crash time cost will be calculated as $L + (i \cdot 2 \cdot 92.11) \cdot D$ where L=normal cost, i=max acceleration, 2=number of hours of work per day, 92.11=overtime rate, and D=number of data scientists on the activity.

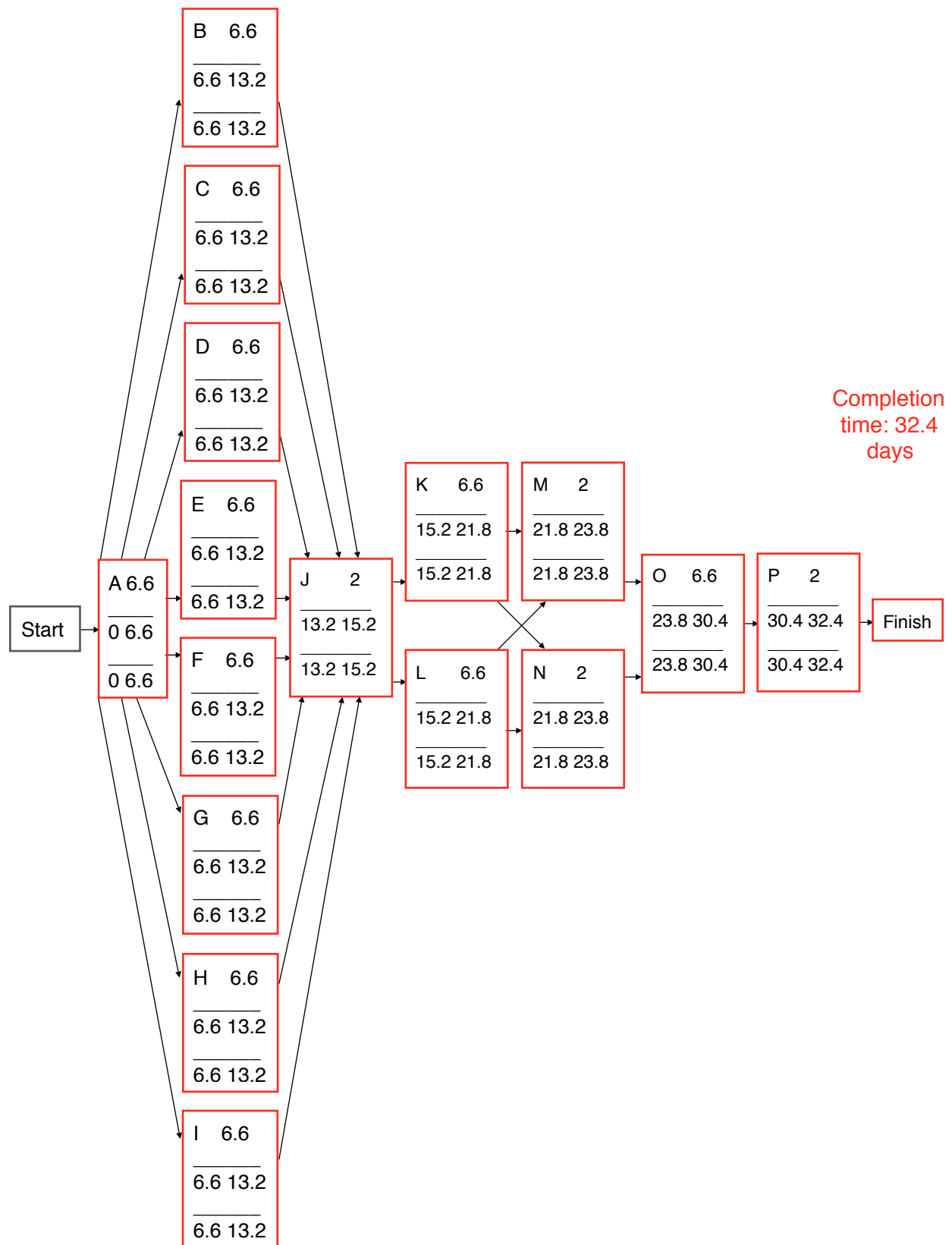
Activity **A** will not be accelerated in order to give enough time to conceptualize and design the project. Given company's guideline on avoiding misuse of overtime work, each activity can not be expedited by more than 2 days. Oriented on these aspects, the table is as follows:

ID	Normal Time (days) "q"	Crash Time (days) "w"	Normal Cost "L"	Crash Cost $[L + (i \cdot 2 \cdot 92.11) \cdot D]$ "e"	Max acceleration (q - w) (days) "i"	Crash Cost Per day $[(e - L) / i]$	On the Critical Path?
A	7	7	6877.92	6877.92	0	n/a	Yes
B	7	5	859.74	1228.18	2	184.22	Yes
C	7	5	859.74	1228.18	2	184.22	Yes
D	7	5	859.74	1228.18	2	184.22	Yes
E	7	5	859.74	1228.18	2	184.22	Yes
F	7	5	859.74	1228.18	2	184.22	Yes
G	7	5	859.74	1228.18	2	184.22	Yes
H	7	5	859.74	1228.18	2	184.22	Yes
I	7	5	859.74	1228.18	2	184.22	Yes
J	2	1	1965.12	3438.88	1	1473.76	Yes
K	7	5	3438.96	4912.72	2	736.88	Yes
L	7	5	3438.96	4912.72	2	736.88	Yes
M	2	1	982.56	1719.44	1	736.88	Yes
N	2	1	982.56	1719.44	1	736.88	Yes
O	7	5	3438.96	4912.72	2	736.88	Yes
P	2	1	982.56	1719.44	1	736.88	Yes
	TOTAL		28985.52				

Activities can be partially crashed.

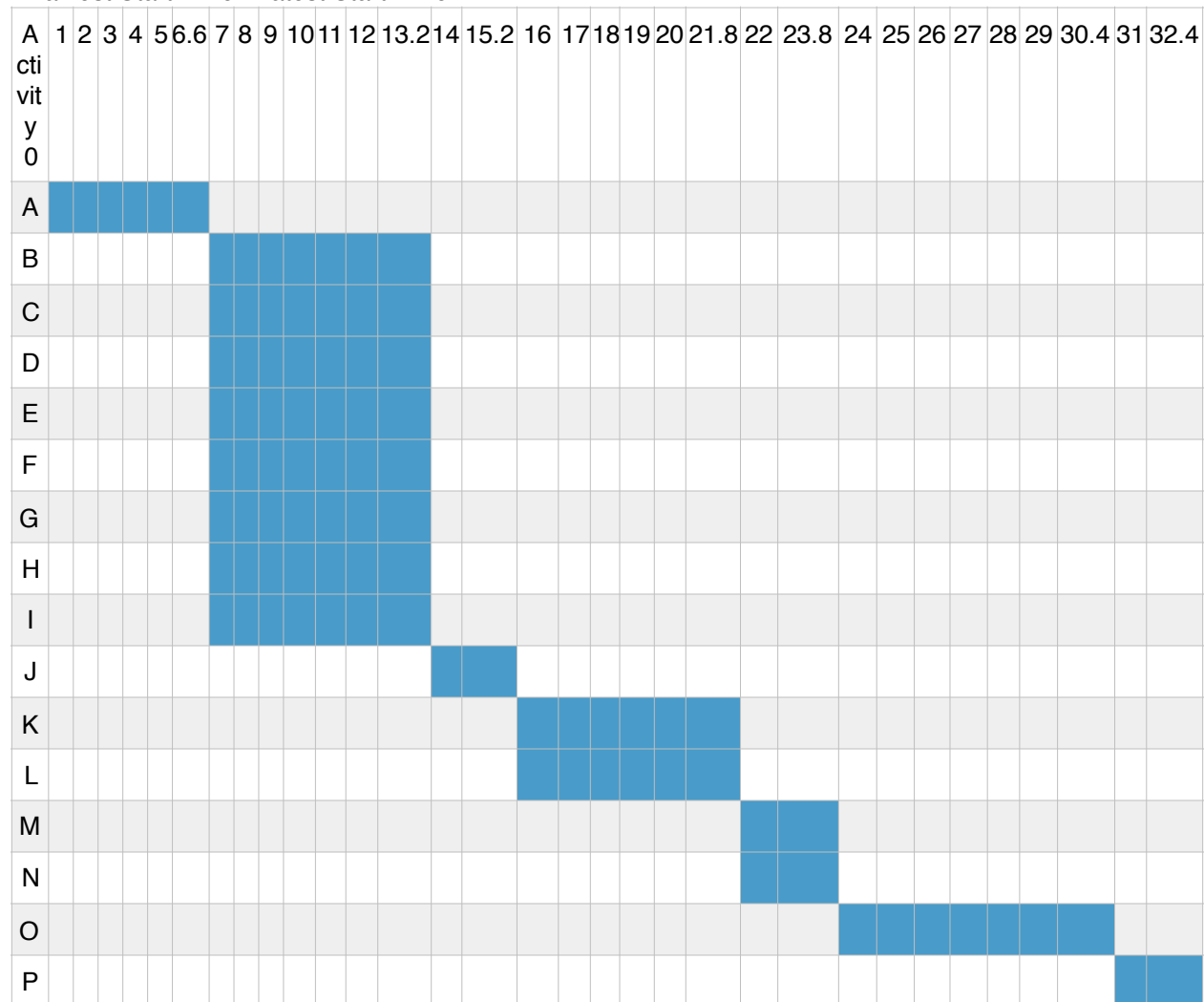
6. Project Scheduling Report

Based on the WBS and the expected time of the activities, the AON network is as follows:



Based on the above AON, a Gantt chart can be built applying the Earliest Start Time and Latest Start Time. Since all activities are critical, one Gantt chart can represent the Earliest Start Time and the Latest Start Time. It is as follows:

- Earliest Start Time / Latest Start Time:



Having the AON as reference, the completion time of the project will be 32.4 days. Also, the total slack(LS-ES or LF-EF) and free slack(smallest ES of successor - EF) are zero for all activities. In addition, all activities are critical, and the critical paths are:

A-B-J-K-M-O-P
 A-C-J-K-M-O-P
 A-D-J-K-M-O-P
 A-E-J-K-M-O-P
 A-F-J-K-M-O-P
 A-G-J-K-M-O-P
 A-H-J-K-M-O-P
 A-I-J-K-M-O-P

A-B-J-K-N-O-P

A-C-J-K-N-O-P
 A-D-J-K-N-O-P
 A-E-J-K-N-O-P
 A-F-J-K-N-O-P
 A-G-J-K-N-O-P
 A-H-J-K-N-O-P
 A-I-J-K-N-O-P

A-B-J-L-M-O-P
 A-C-J-L-M-O-P
 A-D-J-L-M-O-P
 A-E-J-L-M-O-P
 A-F-J-L-M-O-P
 A-G-J-L-M-O-P
 A-H-J-L-M-O-P
 A-I-J-L-M-O-P

A-B-J-L-N-O-P
 A-C-J-L-N-O-P
 A-D-J-L-N-O-P
 A-E-J-L-N-O-P
 A-F-J-L-N-O-P
 A-G-J-L-N-O-P
 A-H-J-L-N-O-P
 A-I-J-L-N-O-P

7. Project Risk Analysis

115% of the expected time, which is 32.4 days, is 37.26 days. In order to compute the probability of the project taking longer than 37.26 days, the variances from the first table on item 5. **Project Activity Analysis** are needed. Since all paths are critical, they have the same completion time, and their variances are the same, any path can be used to calculate the project standard deviation. Taking the critical path A-E-J-K-M-O-P as a reference, the project variance is the sum of the variances of each critical activity on this critical path. As a result, the project variance is $1+1+0.111+1+0+1+0 = 4.111$. Taking the square root of the project variance, the project standard deviation is $\sqrt{4.111} = 2.027$.

The project standard deviation = $\sigma = 2.027$

The expected project completion time = $E = 32.4$

Based on the above values, the Z-value can be computed.

Z-value = $(\text{completion time} - E) / \sigma = (37.26 - 32.4) / 2.027 = 2.397$

Using the Z-value, the respective probability can be found in a Normal Probability Distribution Table. For the Z-value of 2.397 (intersection of 2.4 and 0.00) in the table, the value is 0.9918. As a result, the probability of the project completing within 37.26 days is 99.18%. Given that the idea is to obtain the probability that the project will take longer than 37.26 days, it is needed to subtract from 100% the probability of 99.18%. Therefore, the probability of the project completing longer than 37.26 days is 0.82%, which is a very low probability.

90% of the expected time, which is 32.4 days, is 29.16 days. In order to compute the probability of the project taking less than 29.16 days, the variances from the first table on item 5. **Project Activity Analysis** are needed. Since all paths are critical, they have the same completion time, and their variances are the same, any path can be used to calculate the project standard deviation. Taking the critical path A-E-J-K-M-O-P as a reference, the project variance is the sum of the variances of each critical activity on this critical path. As a result, the project variance is $1+1+0.111+1+0+1+0 = 4.111$. Taking the square root of the project variance, the project standard deviation is $\sqrt{4.111} = 2.027$.

The project standard deviation = $\sigma = 2.027$

The expected project completion time = $E = 32.4$

Based on the above values, the Z-value can be computed.

$$Z\text{-value} = (E - \text{completion time}) / \sigma = (32.4 - 29.16) / 2.027 = 1.598$$

Using the Z-value, the respective probability can be found in a Normal Probability Distribution Table. For the Z-value of 1.598 (intersection of 1.6 and 0.00) in the table, the value is 0.9452. As a result, the probability of the project completing within 29.16 days is 94.52%. Given that the idea is to obtain the probability that the project will take less than 29.16 days, it is needed to subtract from 100% the probability of 94.52%. Therefore, the probability of the project completing in less than 29.16 days is 5.48%.

8. Project Budgeting Report

In order to build a table having the cash flow estimation using early start time and latest start time, the below table is needed. This table contains as columns the activity, ES, LS, E(t), total budget cost per activity, and total budget cost per activity per day based on the Normal Time and Normal Cost columns from the last table on item 5. **Project Activity Analysis**, and the AON on item 6. **Project Scheduling Report**. The table is as follows:

Activity ID	ES(earliest start time) (days)	LS(latest start time) (days)	E(t)(expected completion time) (days) "E"	Total budget cost "B"	Total budget cost per day ("B"/"E")
A	0	0	7	6877.92	982.56
B	6.6	6.6	7	859.74	122.82
C	6.6	6.6	7	859.74	122.82
D	6.6	6.6	7	859.74	122.82
E	6.6	6.6	7	859.74	122.82
F	6.6	6.6	7	859.74	122.82
G	6.6	6.6	7	859.74	122.82
H	6.6	6.6	7	859.74	122.82
I	6.6	6.6	7	859.74	122.82
J	13.2	13.2	2	1965.12	982.56
K	15.2	15.2	7	3438.96	491.28
L	15.2	15.2	7	3438.96	491.28
M	21.8	21.8	2	982.56	491.28
N	21.8	21.8	2	982.56	491.28
O	23.8	23.8	7	3438.96	491.28
P	30.4	30.4	2	982.56	491.28

Since ES=LS, just one cash flow estimation table can represent the daily spending for Early Start Time and Latest Start Time. The table below (split by days among 3 tables) has the cash flow estimation spending on a Gantt Chart format based on the above table:

Activity ID	DAYS														Total
	0	1	2	3	4	5	6.6	7	8	9	10	11	12	13.2	
A		1042.1	1042.1	1042.1	1042.1	1042.1	1042.1								6877.92
B								119.4	119.4	119.4	119.4	119.4	119.4	119.4	859.74
C								119.4	119.4	119.4	119.4	119.4	119.4	119.4	859.74
D								119.4	119.4	119.4	119.4	119.4	119.4	119.4	859.74
E								119.4	119.4	119.4	119.4	119.4	119.4	119.4	859.74
F								119.4	119.4	119.4	119.4	119.4	119.4	119.4	859.74
G								119.4	119.4	119.4	119.4	119.4	119.4	119.4	859.74
H								119.4	119.4	119.4	119.4	119.4	119.4	119.4	859.74
I								119.4	119.4	119.4	119.4	119.4	119.4	119.4	859.74
J															1965.12
K															3438.96
L															3438.96
M															982.56
N															982.56
O															3438.96
P															982.56
Per Day		1042.1	1042.1	1042.1	1042.1	1042.1	1042.1	955.2	955.2	955.2	955.2	955.2	955.2	955.2	
To Date		1042.1	2084.2	3126.3	4168.4	5210.5	6252.6	7207.8	8163.9	9118.2	10073.4	11028.6	11983.8	12939	

Ac tiv ity ID	DAYS								Total
	14	15.2	16	17	18	19	20	21.8	
A									6877.92
B									859.74
C									859.74
D									859.74
E									859.74
F									859.74
G									859.74
H									859.74
I									859.74
J	893.2	893.2							1965.12
K			505.7	505.7	505.7	505.7	505.7	505.7	3438.96
L			505.7	505.7	505.7	505.7	505.7	505.7	3438.96
M									982.56
N									982.56
O									3438.96
P									982.56
P er D ay	893.2	893.2	1011.4	1011.4	1011.4	1011.4	1011.4	1011.4	
To D at e	893.2	1786.4	2797.8	3809.2	4820.6	5832	6843.4	7854.8	

Activity ID	DAYS											Total
	22	23.8	24	25	26	27	28	29	30.4	31	32.4	
A												6877.92
B												859.74
C												859.74
D												859.74
E												859.74
F												859.74
G												859.74
H												859.74
I												859.74
J												1965.12
K												3438.96
L												3438.96
M	350.9	350.9										982.56
N	350.9	350.9										982.56
O			464.7	464.7	464.7	464.7	464.7	464.7	464.7			3438.96
P										409.4	409.4	982.56
Per Day	701.8	701.8	464.7	464.7	464.7	464.7	464.7	464.7	464.7	409.4	409.4	
To Date	701.8	1403.6	1868.3	2333	2797.7	3262.4	3727.1	4191.8	4656.5	5065.9	5475.3	

Even though the budget was increased by 25% as a contingency to address any uncertainty, the intention is to start the project implementation 5 days in advance in order to account for any uncertainty that could result in some need to crash the project. This can minimize extra costs for having 5 extra days to finish the project in case some activity is delayed which could require overtime work to keep the project on schedule. Given that the deliverable date is June 26th 2017, the project will be initiated on May 4th 2017. Also, the cash flow estimation is going to be shared with the accountant people beforehand in order to have them prepared to make the resources available as planned.

9. Project Acceleration Report

Reducing by 15% of its completion time is equivalent to reducing by 4.86. Rounding to the closest full unit, the value will be 5. Based on the last table on item 5. **Project Activity Analysis**, crashing the project by 5 days needs to take into account that all activities are critical, and that activity **A** can not be crashed. The fact that activities **B,C,D,E,F,G,H**, and **I** always need to be crashed altogether makes the cost be **\$1473.76** per unit of time. Therefore, activities **K,L,M,N,O**, and **P** should be crashed first for having the same crashing cost per unit of time and less elements to be crashed for simplicity. While **K** and **L** need to be crashed together, and the same needs to happen with **M** and **N**, **O** and **P** can be crashed individually. Initially, crashing the project by 1 day, using activity **P**, will cost **\$736.88**. This step will increase the project cost from **\$28,985.52** to **\$29,722.40**. Now, crashing the project by 2 days, activity **P** can no longer be crashed. Activity **O** is the one to be crashed this time. Crashing activity **O** by 1 day will cost **\$736.88**. This step will increase the project cost from **\$29,722.40** to **\$30,459.28**. Since activity **O** can be crashed by 2 days in total, it can be crashed one last time. Crashing the project by 3 days, activity **O** will be crashed by 1 day costing **\$736.88**. This step will increase the project cost from **\$30,459.28** to **\$31,196.16**. At this point, crashing the project by 4 days will increase the project cost by the same amount. Therefore, any remaining options can be picked. Choosing activity **J** for simplicity, it can be crashed by only 1 day costing **\$1,473.76**. This step will increase the project cost from **\$31,196.16** to **\$32,669.92**. Finally, crashing the project by 5 days, activities **K** and **L** or **M** and **N** can be the options. Crashing the activities **K** and **L** by 1 day will cost **\$1,473.76**. This step will increase the project cost from **\$32,669.92** to **\$34,143.68**.

Summarizing the whole crashing process:

- crashing the project by 1 day:
 - crashed activity: **P**
 - new total cost of the project: **\$29,722.40**
 - new critical paths(ES=LS): it remains the same critical paths
- crashing the project by 2 days:
 - crashed activities: **P + O**
 - new total cost of the project: **\$30,459.28**
 - new critical paths(ES=LS): it remains the same critical paths
- crashing the project by 3 days:
 - crashed activities: **P + O + O**
 - new total cost of the project: **\$31,196.16**
 - new critical paths(ES=LS): it remains the same critical paths
- crashing the project by 4 days:
 - crashed activities: **P + O + O + J**
 - new total cost of the project: **\$32,669.92**
 - new critical paths(ES=LS): it remains the same critical paths
- crashing the project by 5 days:
 - crashed activities: **P + O + O + J + (K and L)**
 - new total cost of the project: **\$34,143.68**
 - new critical paths(ES=LS): it remains the same critical paths

It is valid to recall that the penultimate table on item 5. **Project Activity Analysis** presents on its last row the total budget increased by 25% as a contingency for any uncertainty that may happen throughout the project development. Remembering this fact, the budget plus a contingency is **\$36,231.90**, which still offers **\$2,088.22** as a remaining contingent value even if the project is crashed by 5 days at a total cost of **\$34,143.68**.