MASTER OF DATA SCIENCE (SEMESTER 2 – 2022/2023)

FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

WQD7007 BIG DATA MANAGEMENT

CASE STUDY

Big Data for Productivity Prediction in the Garment Manufacturing Industry

NAME	YU YUEN HERN
MATRIC NO.	S2121801

Contents

1.0 Big Data Resource

The garment industry exemplifies the industrial globalisation of the modern era. With its labour-intensive nature and manual processes, meeting the global demand for garments catalysed by fast fashion, relies heavily on the productivity of manufacturing employees. Decision makers in this industry highly value tracking, analysing, and predicting team productivity. By leveraging big data and analytics, manufacturers can develop models that consider factors such as historical productivity, production targets, overtime and incentive to accurately predict productivity for optimised resource allocation, targeted training, and efficient strategies to meet the demands of the fast fashion market.

For this, a dataset on the productivity of garment employee has been identified from Kaggle as the big data resource. This dataset contains the relevant features of the garment manufacturing process and employee productivity which were collected from a reputed company in Bangladesh (Imran et al., 2019). Figure 1 details the dataset attributes and description, while Figure 2 features a snapshot of the dataset.

Attribute	Description
date	Date in MM-DD-YYYY
department	Associated department with the instance
team_no	Associated team number with the instance
no_of_workers	Number of workers in each team
no_of_style_change	Number of changes in the style of a particular product
targeted_productivity	Targeted productivity set by the authority for each team for each day
smv	Standard Minute Value, it is the allocated time for a task
wip	Work in progress. Includes the number of unfinished items for products
over_time	Represents the amount of overtime by each team in minutes
incentive	Represents the amount of financial incentive (in BDT) that enables or motivates a particular course of action
idle_time	The amount of time when the production was interrupted due to several reasons
idle_men	The number of workers who were idle due to production interruption
actual_productivity	The actual productivity value which ranges from 0.0 to 1.0

Figure 1: Attributes and description of the dataset (Imran et al., 2019)

date	quarter	departmer	day	team	targeted_p	smv	wip	over_time	incentive	idle_time	idle_men	no_of_styl	no_of_wo	actual_productivity
1/1/2015	Quarter1	sweing	Thursday	8	0.8	26.16	1108	7080	98	0	0	0	59	0.940725
1/1/2015	Quarter1	finishing	Thursday	1	0.75	3.94		960	0	0	0	0	8	0.8865
1/1/2015	Quarter1	sweing	Thursday	11	0.8	11.41	968	3660	50	0	0	0	30.5	0.80057
1/1/2015	Quarter1	sweing	Thursday	12	0.8	11.41	968	3660	50	0	0	0	30.5	0.80057
1/1/2015	Quarter1	sweing	Thursday	6	0.8	25.9	1170	1920	50	0	0	0	56	0.800382
1/1/2015	Quarter1	sweing	Thursday	7	0.8	25.9	984	6720	38	0	0	0	56	0.800125
1/1/2015	Quarter1	finishing	Thursday	2	0.75	3.94		960	0	0	0	0	8	0.755167
1/1/2015	Quarter1	sweing	Thursday	3	0.75	28.08	795	6900	45	0	0	0	57.5	0.753683
1/1/2015	Quarter1	sweing	Thursday	2	0.75	19.87	733	6000	34	0	0	0	55	0.753098
1/1/2015	Quarter1	sweing	Thursday	1	0.75	28.08	681	6900	45	0	0	0	57.5	0.750428
1/1/2015	Quarter1	sweing	Thursday	9	0.7	28.08	872	6900	44	0	0	0	57.5	0.721127
1/1/2015	Quarter1	sweing	Thursday	10	0.75	19.31	578	6480	45	0	0	0	54	0.712205
1/1/2015	Quarter1	sweing	Thursday	5	0.8	11.41	668	3660	50	0	0	0	30.5	0.707046
1/1/2015	Quarter1	finishing	Thursday	10	0.65	3.94		960	0	0	0	0	8	0.705917
1/1/2015	Quarter1	finishing	Thursday	8	0.75	2.9		960	0	0	0	0	8	0.676667
1/1/2015	Quarter1	finishing	Thursday	4	0.75	3.94		2160	0	0	0	0	18	0.593056
1/1/2015	Quarter1	finishing	Thursday	7	0.8	2.9		960	0	0	0	0	8	0.540729
1/1/2015	Quarter1	sweing	Thursday	4	0.65	23.69	861	7200	0	0	0	0	60	0.52118
1/1/2015	Quarter1	finishing	Thursday	11	0.7	4.15		1440	0	0	0	0	12	0.436326
1/3/2015	Quarter1	finishing	Saturday	4	0.8	4.15		6600	0	0	0	0	20	0.988025
1/3/2015	Quarter1	finishing	Saturday	11	0.75	2.9		5640	0	0	0	0	17	0.98788
1/3/2015	Quarter1	finishing	Saturday	9	0.8	4.15		960	0	0	0	0	8	0.956271
1/3/2015	Quarter1	finishing	Saturday	3	0.75	3.94		1560	0	0	0	0	8	0.945278
1/3/2015	Quarter1	finishing	Saturday	1	0.8	3.94		960	0	0	0	0	8	0.902917
1/3/2015	Quarter1	sweing	Saturday	1	0.8	28.08	772	6300	50	0	0	0	56.5	0.800725
1/3/2015	Quarter1	sweing	Saturday	3	0.8	28.08	913	6540	50	0	0	0	54.5	0.800323

Figure 2: A snapshot of the dataset

This dataset fulfils five out of 7V's of big data, namely volume, veracity, visualisation, variability and value. Firstly, the dataset captures a wide range of data points related to worker productivity, such as production metrics, labour inputs, and performance indicators. In addition, each data point is generated by a production line by the end of a work shift. Imagine if a factory has ten production lines, practises two work shifts in a day and runs seven days a week, the factory would have accumulated 7,300 data points by the end of the year. If productivity prediction is proven to enhance competitiveness, garment factories all over Malaysia will adopt this practice and the volume of data would grow. Secondly, the dataset exhibits veracity, as it has been collected directly from garment manufacturing processes, ensuring the accuracy and reliability of the data. Additionally, the dataset can be effectively visualised through charts, graphs, and dashboards. Moreover, the meaning of some attributes in the dataset might change over time. For example, the definition of style change in the *no_of_style_change* attribute could variate in the future as the product design evolves with the market demand. Lastly, the dataset holds significant value for the manufacturing sector, enabling informed decision-making, productivity forecasting, benchmarking, and strategic planning.

This big data resource on the productivity of garment manufacturing workers holds significant value for the manufacturing sector. Firstly, it provides valuable insights into the performance and efficiency of the workforce, allowing manufacturing companies to identify areas of improvement and optimise their operations via better resource allocation, training programs, and process enhancements. Additionally, it can support the development and implementation of predictive models for productivity forecasting, in order to reduce production gap which could cost huge loss to the company (Imran et al., 2019). Moreover, it can serve as a benchmarking tool for manufacturers to compare their productivity performance against industry standards and competitors. This benchmarking analysis can help identify areas where a company may be falling behind or excelling, facilitating a more targeted approach to improvement initiatives.

Alongside the garment manufacturers, the Malaysian Textile Manufacturer Association (MTMA) is the beneficiary of this big data resource. By having concrete data on productivity levels, MTMA can present evidence-based arguments to policymakers and stakeholders, in advocating for the industry's needs and interests, for example, advocating technology adoption to enhance productivity and competitiveness of the Malaysian garment manufacturing industry.

2.0 Storing the Big Data Resource

For storage of this big data resource, the following technologies were considered: MySQL, Hive, Hbase and MongoDB.

To choose the best storage solution suitable for this big data resource, the elimination method was employed. Firstly, the data were collected by the sensor after each shift, transformed by a backend system and then stored in database, therefore the database must be able to handle online transactional processing (OLTP). Hive, despite having similar syntax as SQL, does not possess OLTP capabilities (Özcan et al., 2017), thus it is eliminated. Secondly, as the data will be read most of the time (retrieved for analytical purposes) rather than updated (new records are only added at the end of the shift), read speed is favoured over write speed. Despite MySQL has worse write speed than Hbase, it has better read speed (Bousalem et al., 2019), thus Hbase was eliminated. Thirdly, due to analytical requirements, aggregate operations are crucial for the database. Comparing MySQL and MongoDB, MySQL which is a relational database has better aggregation performance than a document-based NoSQL database like MongoDB (Faraj et al., 2014). Therefore, MongoDB was eliminated and MySQL was chosen as the best candidate for this case study.

3.0 Demo

Now that MySQL is selected as the best candidate for this case study, a prototype demo can be performed. Firstly, MySQL was installed in a Linux environment, then a table was created in MySQL and the dataset in comma-separated values (CSV) format was loaded into the table, as shown in Figure 3, all performed in the MySQL shell.

```
mysql> create table garment (record_date date, quarter varchar(30), department varchar(30), day varchar(30), team int(6), targeted_productivity numeric(10,2), s
mv numeric(10,2), wip int(6), over_time int(6), incentive int(6), idle_time int(6), idle_men int(6), no_of_style_change int(6), no_of_workers numeric(10,2), act
ual_productivity numeric(10,6));
Query OK, 0 rows affected, 7 warnings (0.16 sec)

mysql> show tables;
+-------------------+
| Tables_in_WQD7007 |
+-------------------+
| churn             |
| garment           |
+-------------------+
2 rows in set (0.01 sec)

mysql> select * from garment limit 2;
Empty set (0.01 sec)

mysql> show global variables like 'local_infile';
+---------------+-------+
| Variable_name | Value |
+---------------+-------+
| local_infile  | OFF   |
+---------------+-------+
1 row in set (0.12 sec)

mysql> set global local_infile=1;
Query OK, 0 rows affected (0.00 sec)

mysql> show global variables like 'local_infile';
+---------------+-------+
| Variable_name | Value |
+---------------+-------+
| local_infile  | ON    |
+---------------+-------+
1 row in set (0.00 sec)

mysql> exit;
Bye
student@student-VirtualBox:~$ mysql -uroot -proot --local_infile=1 mysql -e "load data local infile '~/Downloads/garments_worker_productivity.csv' into table WQ
D7007.garment fields terminated by ',' ignore 1 lines"
```

Figure 3: Creating a table and loading dataset into the table using MySQL shell in Linux

From the perspective of the management, it is always good practice to identify the low performing teams or processes in order to find ways to improve them, while rewarding teams which helped the company to be more profitable, oftentimes through cost reduction measures. Table 1 lists down each query and its corresponding output. The reasoning and intention behind each query are as follows, in sequence:

1. To identify the worst three teams to put under performance improvement plans;
2. To identify three tasks which is hardest to perform, thus lower productivity;
3. To identify five most productive teams while having the least workers;
4. To identify five most productive teams while having the least overtime;
5. To identify the teams that does not frequently meet the targeted productivity.

Table 1: Database queries and its corresponding output

<table>
<tr><th>Query</th><th>Output</th></tr>
<tr><td>Find the bottom 3 teams in terms of average actual productivity:
<code>select team, AVG(actual_productivity) as avg_prod from garment group by team order by avg_prod limit 3;</code></td><td><pre>
+------+--------------+
| team | avg_prod |
+------+--------------+
7	0.6680055729
8	0.6741481560
11	0.6819845795
+------+--------------+
3 rows in set (0.03 sec)
</pre></td></tr>
<tr><td>Find the bottom 3 departments (tasks) in terms of average actual productivity:
<code>select department, AVG(actual_productivity) as avg_prod from garment group by department order by avg_prod limit 3;</code></td><td><pre>
+------------+--------------+
| department | avg_prod |
+------------+--------------+
sweing	0.7220130434
finishing	0.7228757108
finishing	0.7820894708
+------------+--------------+
3 rows in set (0.02 sec)
</pre></td></tr>
<tr><td>Find the top 5 teams that has highest actual productivity, sorted ascendingly on average number of workers across all tasks:
<code>select team, AVG(actual_productivity) as avg_prod, avg(no_of_workers) as avg_workers from garment group by team order by avg_prod desc, avg_workers limit 5;</code></td><td><pre>
+------+--------------+-------------+
| team | avg_prod | avg_workers |
+------+--------------+-------------+
1	0.8210543619	35.042857
3	0.8038798316	39.521053
12	0.7790554444	23.919192
2	0.7708551376	34.623853
4	0.7700348476	38.200000
+------+--------------+-------------+
5 rows in set (0.02 sec)
</pre></td></tr>
<tr><td>Find the top 5 teams that has highest actual productivity, sorted ascendingly on sum of overtime:
<code>select team, AVG(actual_productivity) as avg_prod, SUM(no_of_workers) as sum_overtime from garment group by team order by avg_prod desc, sum_overtime limit 5;</code></td><td><pre>
+------+--------------+--------------+
| team | avg_prod | sum_overtime |
+------+--------------+--------------+
1	0.8210543619	3679.50
3	0.8038798316	3754.50
12	0.7790554444	2368.00
2	0.7708551376	3774.00
4	0.7700348476	4011.00
+------+--------------+--------------+
5 rows in set (0.02 sec)
</pre></td></tr>
<tr><td>Find the top 5 teams that has highest difference between average targeted productivity and average actual productivity:
<code>select team, AVG(actual_productivity)-AVG(targeted_productivity) as prod_diff from garment group by team order by prod_diff desc limit 5;</code></td><td><pre>
+------+--------------+
| team | prod_diff |
+------+--------------+
1	0.0743876959
3	0.0617745686
4	0.0524158006
2	0.0309468816
5	0.0243248719
+------+--------------+
5 rows in set (0.02 sec)
</pre></td></tr>
</table>

4.0 Big Data Pipeline

According to Imran et al. (2019), one common problem faced by garment manufacturers is the productivity gap of the employees, which occur when the actual productivity does not meet the targeted productivity, thus causing the company to face huge loss. This proposed big data pipeline as illustrated in Figure 4 aims to solve this problem by allowing decision makers to input data and then being served with predictive analytics and visualisation.

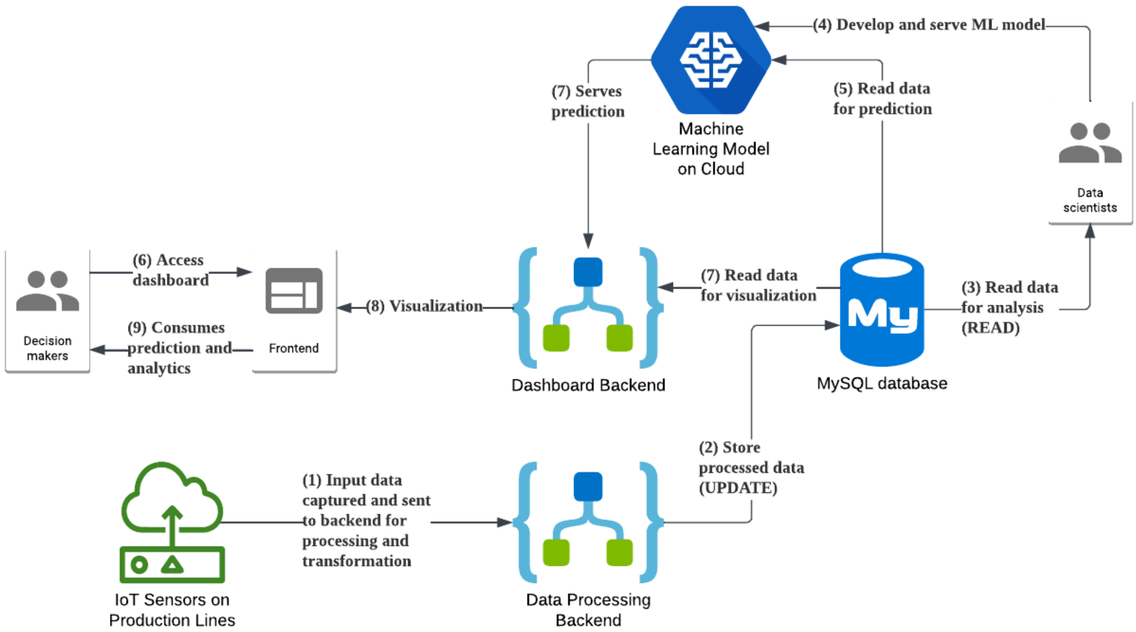

Figure 4: Proposed big data pipeline

The data pipeline illustrated above can be related to the six phases of big data as follows:

1. Data Generation: The initial phase involves sensors on the production lines, which generates data that is further processed.
2. Data Acquisition: The sensor input data is sent from the sensors to the data processing backend. The data processing backend receives the data, performs transformation and stores it in a MySQL database via an update operation.
3. Data Storage: The data stored in the MySQL database represents the data storage phase. The backend updates the transformed sensor data in the database, making it available for future retrieval and analysis.
4. Data Analysis: A group of data scientists reads the data from the MySQL database for analysis and development of the predictive model. This step involves exploratory data analysis, model development and evaluation.
5. Data Visualisation: Once the machine learning model is developed, it is served on the cloud. When a user requests a prediction, the backend retrieves the necessary data from the database and makes a call to the machine learning model. The retrieved data is then

visualised on the frontend together with the predictions, enabling users to have a visual representation for better understanding and decision-making.

6. Decision Making: Users consume the predictions and visualisation presented on the frontend, which leads to the decision-making phase. They make data-driven decisions based on the insights and predictions derived from the data.

5.0 Data Provenance

Data provenance is critical in the garment manufacturing use case. It ensures transparency and traceability of the data throughout its journey, from the origin of the sensor data to the development of predictive models, thus providing a trail of information that can be leveraged to address data inaccuracy issues. Data provenance mechanisms are implemented in the data processing backend which captures details such as sensor types, calibration information, timestamps, and production lines and machineries associated with the data, subsequently stored in a separate table in the MySQL database.

In addition to data provenance on raw sensor data, the entire big data pipeline will undergo data provenance procedures as well. Figure 5 illustrates the directed acrylic graph (DAG) for all processes involved. As the entire process has been verbosely explained in Section 4.0, therefore the DAG will not be further explained.

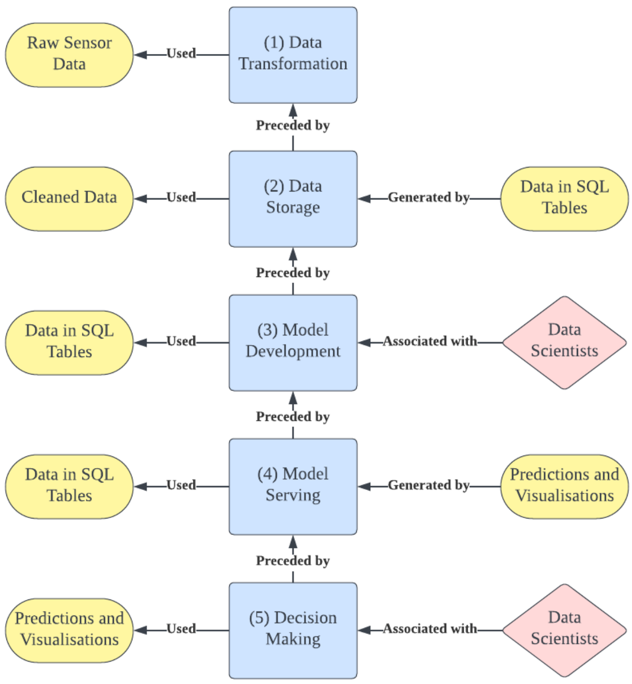

Figure 5: Directed acrylic graph of all processes involved

When unexpected data inaccuracy occurs, stakeholders can utilise the lineage information about the data to trace back to the root cause of the inaccurate data. In the context of classification models, source data needs to be assessed and validated again if the model kept predicting the same class despite the features being distinctively different. By examining the provenance information, stakeholders can identify potential causes such as sensor malfunctions, calibration errors, data collection anomalies or even data transformation mistakes. This allows for targeted investigations and corrective measures to mitigate or prevent similar inaccuracies in the future.

References

Bousalem, Z., Guabassi, I. E., & Cherti, I. (2019). Relational Databases Versus HBase: An Experimental Evaluation. *Advances in Science, Technology and Engineering Systems Journal*, *4*(2), 395–401. https://doi.org/10.25046/aj040249

Faraj, A., Rashid, B., & Shareef, T. (2014). *COMPARATIVE STUDY OF RELATIONAL AND NON- RELATIONS DATABASE PERFORMANCES USING ORACLE AND MONGODB SYSTEMS*. *5*(11).

Imran, A. A., Amin, M. N., Islam Rifat, M. R., & Mehreen, S. (2019). Deep Neural Network Approach for Predicting the Productivity of Garment Employees. *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*, 1402–1407. https://doi.org/10.1109/CoDIT.2019.8820486

Özcan, F., Tian, Y., & Tözün, P. (2017). Hybrid Transactional/Analytical Processing: A Survey. *Proceedings of the 2017 ACM International Conference on Management of Data*, 1771–1775. https://doi.org/10.1145/3035918.3054784