**5240 - Harvesting, Storing and Retrieving Data - Final Project**

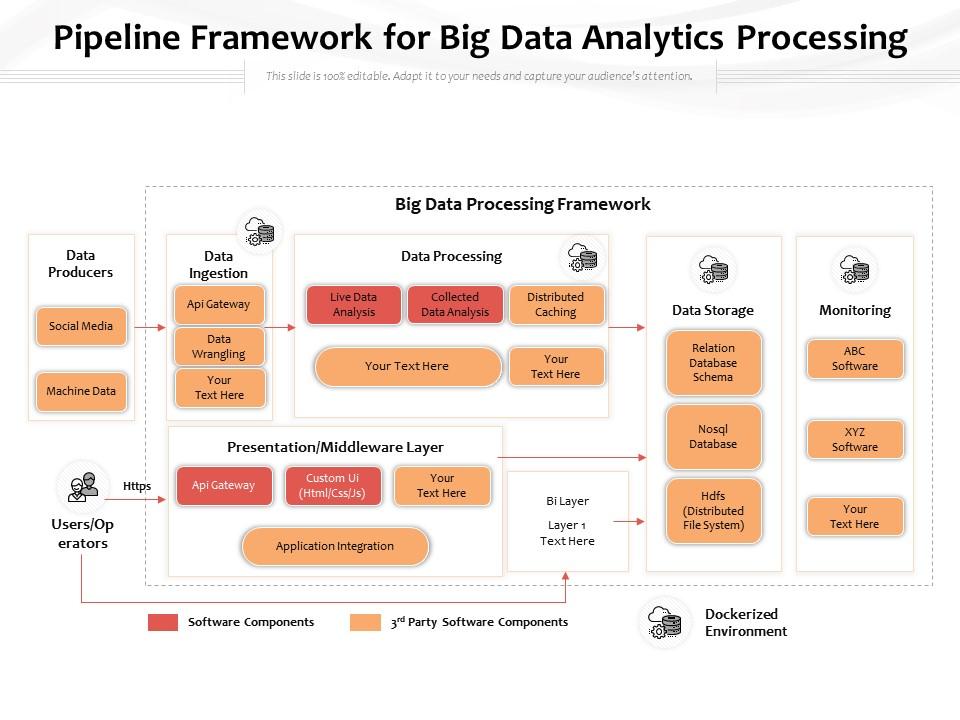
**Introduction**

The era of digital transformation has shown how traditional data processing models are no longer enough to cope with the high volume, velocity and variety of the data being generated. An organization’s success depends on the insights it gathers from the data it generates and the potential of big data analytics allows to unlock the valuable insights hidden within the raw big data. Through this project we aim to implement a big data model that enables us to process and analyze a large dataset with diverse feature.

**Part 1: A Strategy to Deploy Big Data Technology in a Firm**

In this data driven world, it’s essential for any company to take leverage of the power of big data technology to enhance its market position amidst aggressive competition. The deployment of big data technology represents a strategic investment for our future growth. By utilizing the data analytics techniques, we could deliver valuable insights, decision making which would be data-driven and deliver the improved products and services to our customers. Being the Chief Information/Data Officer (CIO/CDO) of a publicly listed company, I would like to present the following strategy to deploy big data application company-wide to cater board’s vision to gain an edge over the competitors and to achieve the business goals. The strategy in question employs various aspects of the firm, including required technology selection, human capital development and the measures to aggressive counter the competition. We hold a better position in the market, however, to maintain the same and improve our position with growing competitors, it is most necessary to harness the power of big data.

A framework is required to select the right data deployment model. Below flowchart provides one of the best deployment options.



**Flowchart: Big Data Processing Framework** (Pipeline framework for big data analytics processing, n.d.)

**Assessment of Firm's Characteristics:** Before coming up with a deployment strategy, it's necessary to conduct a viable assessment of the firm's size, sector, and competitive environment:

* **Size:** Knowing the scale at which the organization operates, including the volume of data it generates and the resources available.
* **Sector:** Understanding the dynamic industry regulations, potential challenges, and the opportunities within the sector.
* **Competitors:** Analyzing competitors approaches to employing big data technology and identifying areas for competitive advantage.

**Assessment of Competitors:** Most of the competitors within the sector are engaging with advanced technologies to streamline the daily operations, improving customer experience, and driving innovation. To remain competitive in this dynamic industry, a proactive approach needs to be taken towards technological advancements.

**Technology Selection:** Considering the various needs for any business and the diverse operations they perform, the big data technology ecosystem in consideration should offer robust analytical capabilities. The proposal of the following big data deployment model and its components allows improving the requirements of any organization.

* **Google Cloud Platform (GCP):** Cloud technology offers a suite of services that help with big data deployment, it allows for quick scaling the capacity and be flexible with the pricing needs. The AI platforms being integrated into the cloud services allow us to analyze large volumes of data at times without human involvement. The pay as you require pricing model allows it to provide a cost-effective solution to quickly scale infrastructure and allow companies to improve their efficiency and accelerate their decision-making process.
* **OpenRefine for Data Cleaning and Transformation:** Open refine was formerly called as Google Refine, this helps with handling large datasets of different types and provides way with cleaning data and tasks that include transformations, slew of operations that are present within the tool allows to implement powerful data wrangling methods as well and its user-friendly interface allows to prepare data for further analysis at ease.
* **BigQuery on Google Cloud Platform (GCP):** BigQuery's managed data warehouse service allows us to store and analyze large amounts of data with greater efficiency. The serverless architecture and its thoughtful integration with other GCP services makes it a great choice for the project’s requirements. The data can be quickly imported as tables with suitable schema and operation can be performed with the help of queries.

**Human Capital Development:** The successful implementation of any big data technology totally relies on the skills and competence of any organization’s workforce. To ensure that the staff are equipped with the necessary expertise, training and seminars need to be conducted this allows anyone to get familiarizes with the implementing big data technologies, in addition to upskilling the existing workforce efforts need to be made to recruit specialized talent who have diverse expertise in big data.

**Part 2 : Data Preprocessing Steps**

Company A runs its business in all the US regions (East, West, Central, and South), and it has its operation in almost all the states. They have been doing business to serve three segments of customers: Consumer, Corporate, and Home Office. The company sold products in three categories: furniture, office supplies, and technologies. Each category is divided into several sub-categories.

In the [dataset](https://unt.instructure.com/courses/102892/quizzes/592989/$CANVAS_OBJECT_REFERENCE$/modules/gdc5436dcb2522ffc675c7d19bca38ca8) [**sales\_record\_1\_ADTA5240.csv**](https://unt.instructure.com/courses/102892/files/25211328/download?wrap=1), in each record (a sale transaction), the company collected data of the customer ID, the sale amount of the transaction, the zip code, and the state where the transaction occurs, and all other related data.

The data file a CSV file in which each field in a record is separated from the next ones by commas. It means the separator used in the CSV file is a comma. Among the attributes (variables or columns) of the dataset is Product Name, of which the values may contain some comma(s).

The comma ‘,’ is used as the separator in the CSV data file. Therefore, the existence of commas ‘,’ in the values of the attribute (or column) “Product Name” may cause problems when some query is run on the data set. **So, data must be pre-processed before being used, the following describes in detail of all the steps that has been taken to preprocess the dataset.**

1. **OpenRefine:**

OpenRefine can be accessed through web browser or desktop application.

A screenshot of a computer

Description automatically generated

1. **Importing the Dataset:**
   * Click on “Create Project” and choose “Get data from This Computer”.
   * Navigate to the location of the CSV file “sales\_record\_1\_ADTA5240.csv” and select it.

A close up of a text

Description automatically generated

* + Click “Next” to import the dataset into OpenRefine.
  + Click “Create project” to complete the Import.

A screenshot of a computer

Description automatically generated

1. **Find and Replace:**

* In the OpenRefine workspace, locate the “Product Name” column in the dataset.



* Click on the drop-down arrow next to the “Product Name” column.
* Select “Edit cells” -> “Replace”.

A screenshot of a computer

Description automatically generated

* In the “Find” field, enter a comma (,).
* In the “Replace with” field, enter a hyphen (-).
* Click “OK” to replace all commas with hyphens in the “Product Name” column.

A screenshot of a computer

Description automatically generated



1. **Verification:**

Scroll through the "Product Name" column to ensure that all commas have been replaced with hyphens, or we can visually inspect using the Facet-> Text facet as well.

A close-up of a sign

Description automatically generated **converted into** A close up of a sign

Description automatically generated

1. **Export the Pre-processed Dataset:**

* Click on the "Export" button in the top-right corner.



* Choose the “Comma-separated value” format for exporting the dataset.
* The file will be downloaded “sales-records-1-ADTA5240-csv.csv”. Renamed the csv as per the requirement below.



**Part 3: Verifying the Cleaned Dataset:**

Needs to check and verify that the data about the state names and zip codes are correct. If any error (typo) is found, the error should be corrected before the data set is used in the analysis. Required to use OpenRefine again to perform the check and correct the error, including doing it manually.

1. **State Names:**

* Using the text facet option in the drop-down option next to the “State” column, we can check on the possible errors.

A screenshot of a computer

Description automatically generated

* When using the “Cluster” option in the text facet the following errors can be seen.

A screenshot of a computer

Description automatically generated

* The Method “Key collision” and Keying function “Methaphone3” was employed.
* The errors are for state of Florida and Nebraska.
* So, checking the “Merge?” for both the “Values in cluster” and then click “Merge selected & re-cluster” to change the values as per the “New cell value”.
* After the change below changes can be seen, for example the “Nebreska” in row 23 is changed to “Nebraska”.



After change

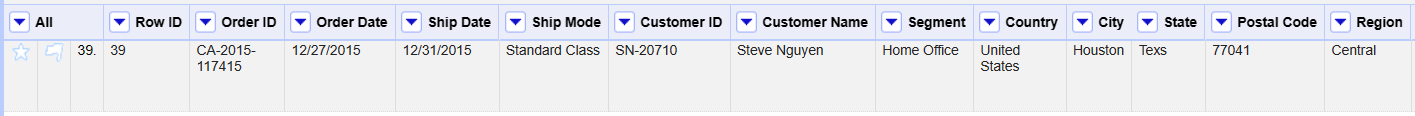


* Similarly, the same has been changed for Florida as well.
* To re-verify the errors, the facet is re-visited, and the below error can be seen for the state of Texas.

A screenshot of a computer

Description automatically generated

* The above can be edited manually by clicking on “Texs” row in the facet as it includes only one row.



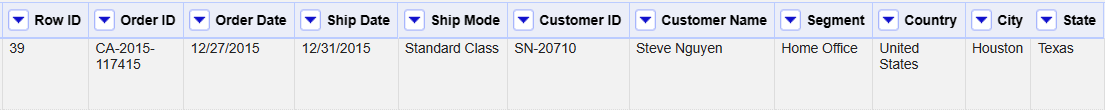
* Click on “edit” next to the “Texs” and change it manually.

A screenshot of a cell phone

Description automatically generated

* Make the change and click on “Apply” . The following pop-up can be seen with the changes applied.





* Checking the text facet again, shows there is only one choice for Texas.

A screenshot of a computer

Description automatically generated

1. **Postal Code:**

The postal code in US would be of usually 5-digit number or at time’s a 5-digit code followed by 4-digit postal code. So, need to check if there are any other unique values.

* The above can be achieved using the text facet option. Click on the drop-down arrow next to the column header containing Postal codes. Select "Facet" > "Text facet" to create a facet for the Postal codes.

***Note:*** *Using “Text facet” but not “Numeric Facet” as there as all the Postal codes are as text but not as numeric values.*

A screenshot of a computer

Description automatically generated

* The facet can be visually inspected and there are choices less than 4-digits, for ex. 1040.
* These 4-digits can be filtered using the “Custom text facet” under the facet option of the column Postal codes. Enter the Expression “value.length()<5” and click “OK”.

A screenshot of a computer

Description automatically generated

* The below results can be seen for the custom facet of Post code. Where there are 438 values that are of 4-digit or lesser.

A screenshot of a computer code

Description automatically generated

* Clicking on the “true” will show the below in the “Text facet”. Here all the 438 rows are listed with 4-digit code.

A screenshot of a computer

Description automatically generated

* To mitigate the above a “0” can be added to the start of the value. The following expression **if(value.length()<5, “0”+value, value)** can be used in the Edit cells -> Transform. This can be used to update the all the 5-digit to 4-digit values.

A screenshot of a computer

Description automatically generated

* To verify, the “Custom text facet” with expression can be used again.

A screenshot of a computer

Description automatically generated

* This shows that the there are no choices below 5-digit.

A screenshot of a computer code

Description automatically generated

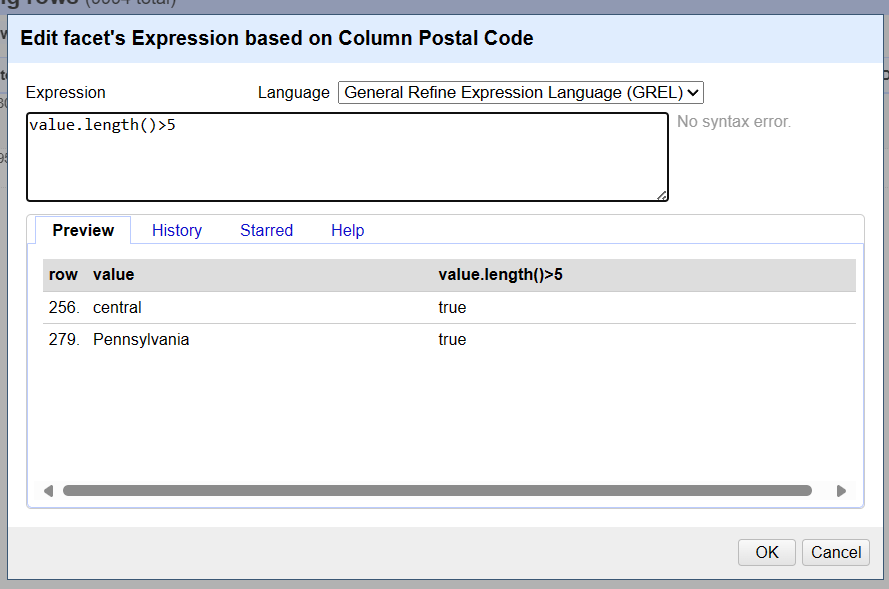
And all the 4-digit values are appended at the start with “0”.

A screenshot of a computer

Description automatically generated

* Checking if there are any values greater than 5-digit. As all of these are text and not numeric, using the “Custom text facet” again on the “Postal code” column with expression,

**“Value.length()>5”** this results in two rows with text rather than postal codes.



* The below two rows are the one with error values and these need to be updated manually.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* Filtering the “State” column “Illinois” and the value of 256th row has error, this can be updated with the value of above column as it has the same location details.

A white paper with black text

Description automatically generated

* Click on the “edit” option next to the “central” and update the value with “60623”.

A screenshot of a computer

Description automatically generated

* After changes the below can be seen in the 256th row.

A white background with black text

Description automatically generated

* Similarly, the value of “Pennsylvania” in the 279th row need to be edited manually.

A white paper with black text

Description automatically generated

* Click on the “edit” option next to the “Pennsylvania” and update the value with “19140”.

A screenshot of a computer

Description automatically generated

* After changes the below can be seen in the 279th row.

A white page with black text

Description automatically generated

* With this all the values are updated as 5-digits, including all the text values are converted to 5-digit postal codes.
* Exporting the dataset as “Comma-separated values” and saving the cleaned dataset as **“Cleaned\_records\_3.csv”.**

****

**Part 4: Big Query**

The SVP (Senior Vice President) of Business Strategy wants to design a new business strategy based on the business insights that can be extracted from the collected dataset (It is also assumed that the student is his/her Senior Assistant). The SVP asked the student to provide the following pieces of information that can be queried from the dataset:

(a) The top five states where the company has the most number of customers.

(b) The top ten zip-code areas where the company gets the best sale amounts, i.e., the highest sale figures.

**Importing dataset in GCP:**

* To create a dataset in the cloud console to store the data, go to “Navigation menu -> APIs and services ->Enabled APIs and services”, then search for the BigQuery API and Enable if it’s not enabled already.

***Note****: Working under the project ID: adta5240sarvepalli*

* Go to “BigQuery -> BigQuery Studio”

A screenshot of a computer

Description automatically generated

* In the above seen dashboard, clicking on the three vertical dots of the project **adta5240sarvepalli,** this shows the option to **Create data set,** then clicking on it to create the dataset.

A screenshot of a computer

Description automatically generated

The below pop-up appears and updating the following information,

Data set ID – Cleaned\_records

Location type – Multi-region

Multi-region – US (multiple regions in United States)

Advanced options – Leaving them as default

A screenshot of a computer

Description automatically generated

Clicking on the shows the following dataset created under the Project ID.

A screenshot of a search engine

Description automatically generated

* The dataset info can be viewed by clicking on the “Cleaned\_records”, it also shows an option to create a table.

A screenshot of a computer

Description automatically generated

* Clicking on “CREATE TABLE” open up a pop-up and the below details are filled to create a Table.

Create table from – Upload

Select file – Cleaned\_records\_3.csv

File format – CSV

Project – adta5240sarvepalli

Data set – Cleaned\_records

Table – Cleaned\_records3

Schema – Used the below details,

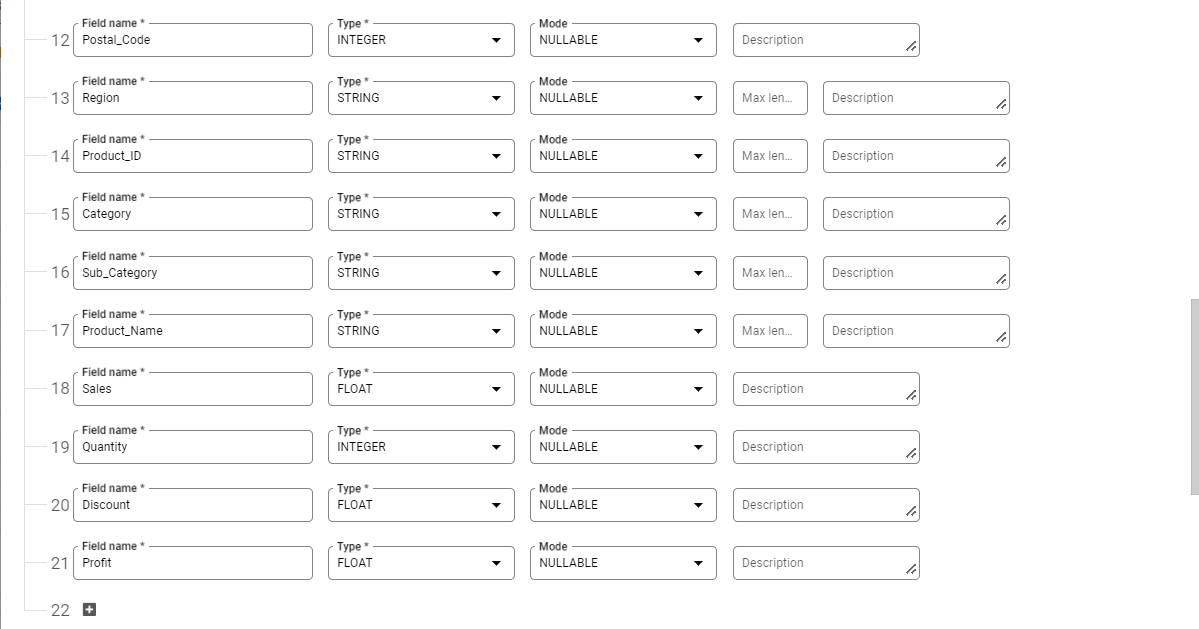
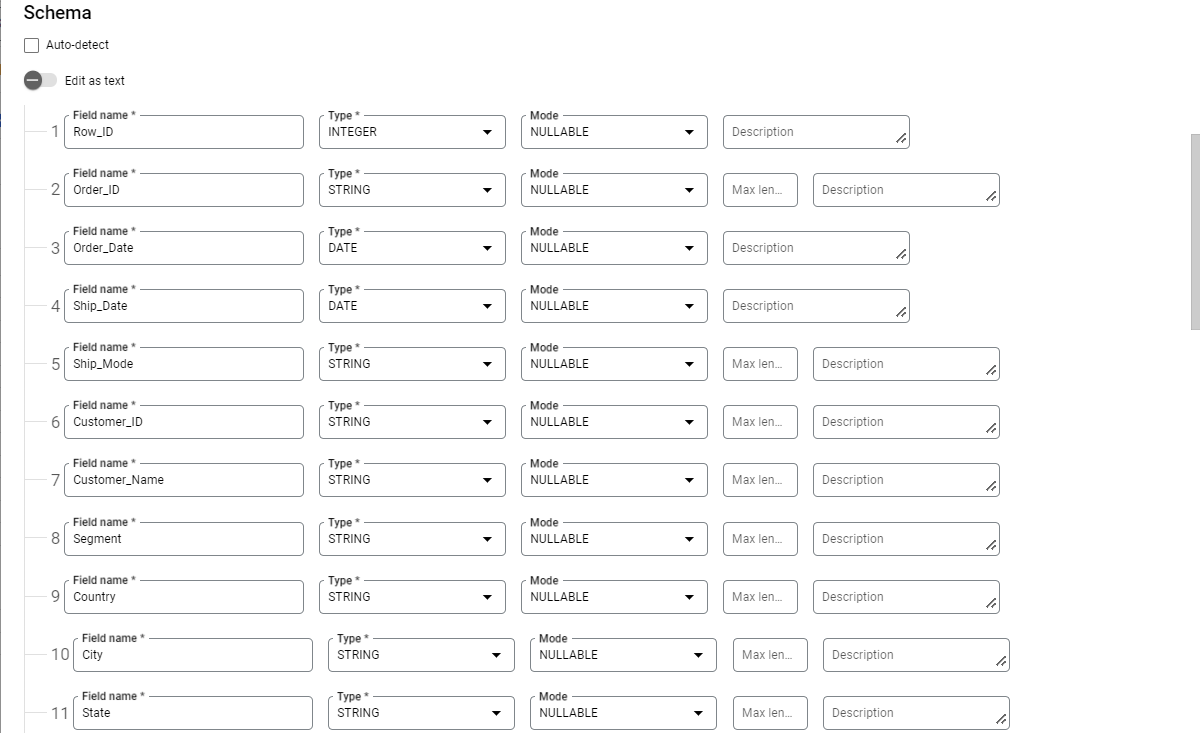
*Row ID:integer, Order\_ID:string, Order\_Date:string,Ship\_Date:string, Ship\_Mode:integer, Customer\_ID:string, Customer\_Name:string, Segment:string, Country: string, City: string, State:integer, Postal\_Code:integer, Region: string, Product\_ID: string, Category: string, Sub\_Category: string, Product\_Name: string, Sales:float, Quantity: integer, Discount:integer, Profit:float*

In the Advanced options,

Field delimiter - Comma

Header rows to skip - 1

A white sheet with black lines

Description automatically generatedA screenshot of a computer

Description automatically generated

* The Table “Cleaned\_Records3” will be created and the Schema details can be seen in the right pane.

A screenshot of a computer

Description automatically generated

* The table can be previewed, which shows the whole cleaned data.

A screenshot of a computer

Description automatically generated

* In order query the table, “create SQL query” can be used to input the Query as per the requirement.

A screenshot of a computer

Description automatically generated

* The Query for, **The top five states where the company has the most number of customers.**

A screenshot of a computer

Description automatically generated

* The Query results shows the top 5 states with most number of customers.

A screenshot of a results report

Description automatically generated

* A new Query editor can be opened and the query for, **The top ten zip-code areas where the company gets the best sale amounts, i.e., the highest sale figures**

A screenshot of a computer

Description automatically generated

* The Query results for the top-10 Postal Codes and their associated Total Sales amount.

A screenshot of a computer

Description automatically generated

**Conclusion:**

In conclusion, The project showed how various aspects of big data technology can be deployed in a firm, the data preprocessing, the verification, and querying of the cleaned dataset using BigQuery.

The firm’s size, sector and competitive landscape was considered while designing the plan. The components for the ecosystem were identified and the GCP BigQuery, OpenRefine were employed to complete the project.

The Data preprocessing played a very serious role to ensure that the data’s quality and consistency is maintained. By replacing the commas in the “Product Name” column and also verifying the “State Names” as well as the “Postal codes”, thus making the cleaned dataset fit for analysis.

Valuable insights were drawn from the results of BigQuery, this includes the top states with the most customers and the top Postal codes with the most sale amounts. This kind of insightful data is useful to make strategic decisions and drive business growth.

To summarize, the projects showed the importance of leveraging big data technologies, tedious data preprocessing and robust analysis to have a competitive edge in achieving the business objectives, in the ever-improving business landscape.

# References

*Pipeline framework for big data analytics processing*. (n.d.). Retrieved from Slide Team: https://www.slideteam.net/pipeline-framework-for-big-data-analytics-processing.html