**My Store Business Analysis -Batch Edition.**

INTRODUCTION **:**

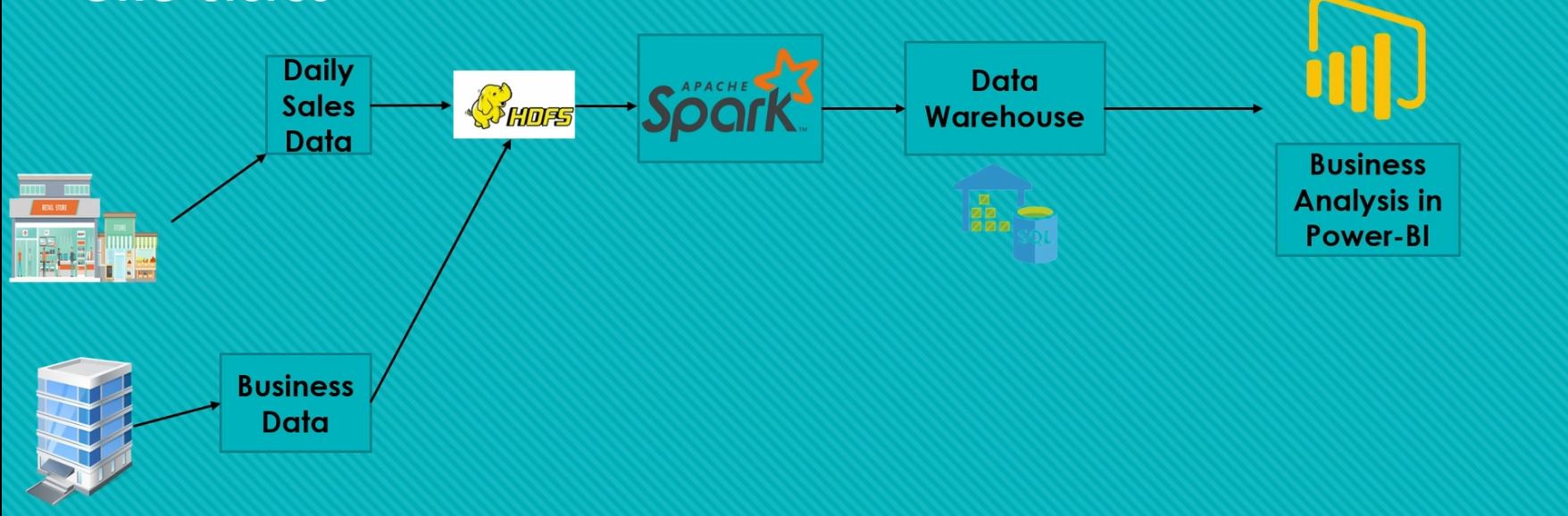
* In this project we build hypothetical company or business that needs Big data solutions to be implement for their business uses.
* The company has provided use cases that we will understanding, planning, implementing in our project.
* My-Store is a chain of retail stores, serving the customers world-wide, with a wide range of products like groceries home decor, furniture, electronics etc.
* My-Stores has 100+ retails stores, selling worldwide and generating the sales data at their POS (point of Sale) systems.

**Tech Stack:**

* Operating System -Windows/Linux,
* Processing Framework - Hadoop HDFS, Apache Spark,
* Local integration for code environment (IntelliJ Idea)
* Datawarehouse- MySQL
* Scheduling Workflow -Apache Airflow
* Dashboard Visualization -Power BI

**Pipeline Design:**

* As we can see, we have a huge number of stores that will be pushing the daily sales data from their online server into our big data pipeline. Once we get the daily data, we will be ingesting this into our HDFS Hadoop distributed file system. Along with that, we will also have to consider some business data directly from the My stores headquarters, which might have many useful business data that we have to consider while processing the daily sales.
* We will also read that into our HDFS.
* Once all the data is inside our HDFS, we will use Apache Spark, the distributed computing platform to process this huge amount of data effectively and efficiently. Then post processing, we will aggregate all the data and store it in our data warehouse. Then from the data warehouse we will be pulling that data for business analysis using Power Bi as reporting and visualization tool along with that as this is the batch data, we will be scheduling all this pipeline using Apache Airflow so that the pipeline remains consistent even in cases of failure.



**Business Requirements:**

Business says that the data that we will be getting in our pipeline will be in the form of sale ID, product ID, Quantity sold, Vendor ID, Sale date, Sale amount, sale currency, and they have kept few conditions as the data is being generated on the single POS systems connected to the network.

**Columns in Daily Data :**

Sale\_ID, Product\_ID, Quantity\_sold, Vendor\_ID, Sale\_date, Sale\_amount, Sale\_currency

**Conditions:**

* Due to few network issues, there might be few columns that can be or cannot be missing.
* If Missing Quantity\_sold - Hold the data till the updates.
* Vendor\_ID - Hold the data till the updates.
* Release the hold records in case updates of the missing values is received in consecutive day’s data.
* Sale\_amount -Derive from the product price, provided by the Headquarters, considering Sale Currency’s exchange rate in USD.
* Get the Vendor details from the Business data, and bind with the final extract, for further reporting and analysis.
* Get the complete sales data, per Sale\_ID, to be fetched into the report.

**Deployment :**

For the deploy the code u need to installed above mentioned tech stack in your machine then Linux terminal inside in your project target location file

ex. Cd MyStore\_BusinessAnalysis\Code\MyStoreproject\target> Run below commands

* 'spark-submit --master -yarn --deploy -mode cluster --jars config-1.2.1.jar --class DailyDataIngestAndRefine MyStoreproject-1.0-SNAPSHOT.jar'
* spark-submit --master -yarn --deploy -mode cluster --jars config-1.2.1.jar --class EnrichProductReference MyStoreproject-1.0-SNAPSHOT.jar
* spark-submit --master -yarn --deploy -mode cluster --jars config-1.2.1.jar --class VendorEnrichment MyStoreproject-1.0-SNAPSHOT.jar