ProPharma – Data Store & Operational Analytics

Purpose of the Data Store & Operational Analytics

- To handle, manage and visualize Data, related to day-to-day operational activities

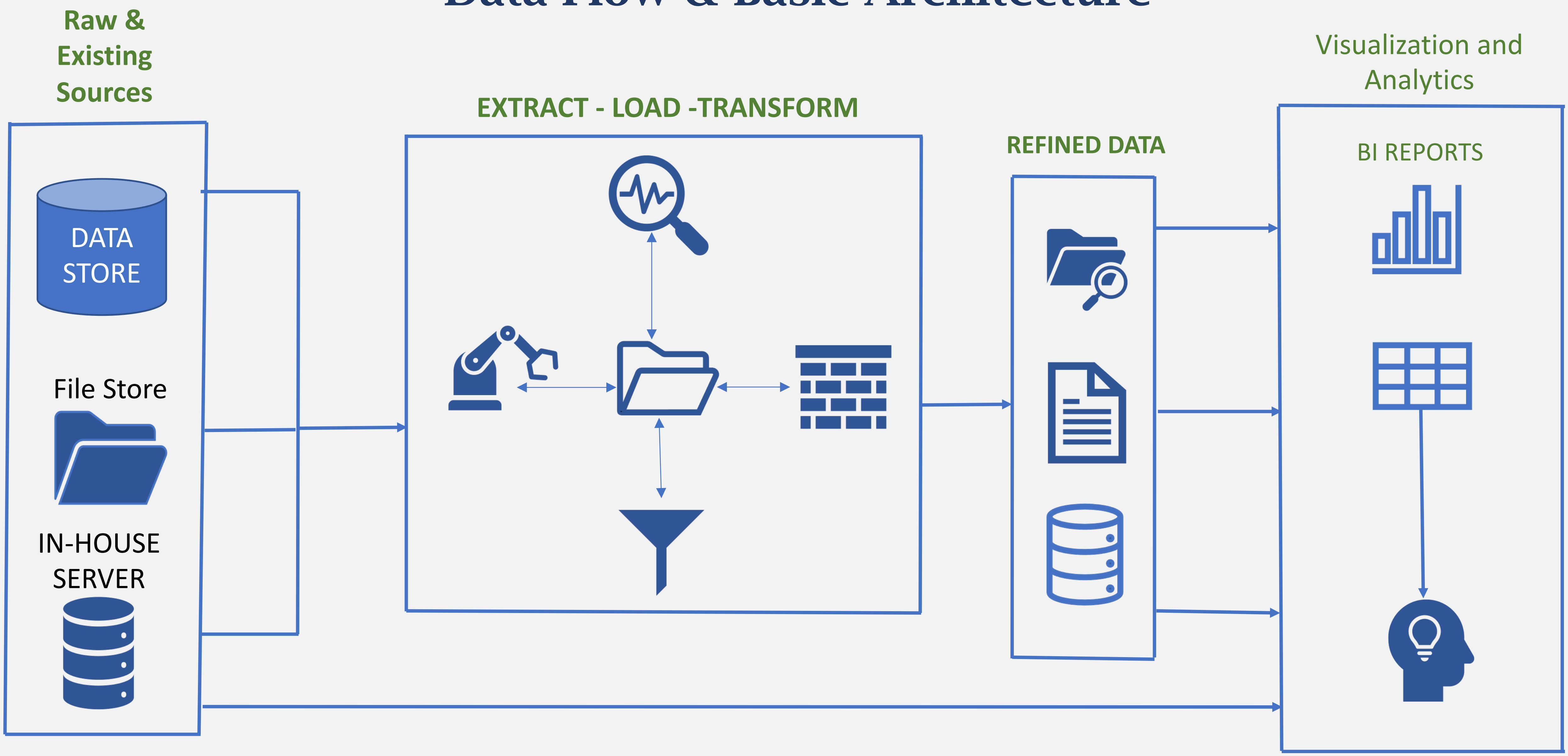
- Collate Pull and Capture RAW data from **multiple Sources** – like On-Premise, cloud, Social Media(like Twitter),Share Drive Data, Web Forms either(API or Curl script) etc.,

- Apply **Extract and Load process** on the collated data and push it into Data lake location – keep it as one source of trusted data source for further data processing and refinement

- Apply a relevant script (like python programming) model on the incoming data to Data lake and **Transform and Load it again back to Data Lake and systematically into Data base /Data Warehouse.**

- The refined and transformed data in the Data Base or Data Warehouse can we used Reporting and Machine learning

Data Flow & Basic Architecture



Infrastructure/Environment Selection : Cloud , Why?

- Readily available resources
- No license fees for cloud services, No Environmental readiness is required and No upfront installation costs
- Data can be replicated across multiple regions to help fall back and Failure overs
- We can have read replication to help the Data application server better
- Scalability according to data growth is easier and require no extra commitments.
- Cost applicable only for services based on the commute power, execution load and run time
- Availability all the resources , ETL , Data Lake, Data transformation engine, DB/DWH , Reporting and Machine learning tools in single space.

Data Sources

Data Sources to be considered and not limited to the following below.

General Dataset to ProPharma

1. Country/Regional Demographic information
2. Distributors and their Performance
3. Sales Representatives and Performance
4. Drugs sold out – Day-to-Day Sales
5. Corporates/companies buying new drugs
6. Other Market competitors with similar drugs

Relevant only to New Drug launched by Propharma

1. Clinical Trials
2. Pricing of the Medicine
3. Patient Testimonials
4. Adverse Events
5. Sales Operations and CRM Software
   - Collate data about customer relationships.
   - Documentation of sales team performance.
   - Record customer's decision-making process.
   - operational costs for sales operations and CRM.

5. Social Media – (like : Twitter, Quora)
6. Collating Data from Pharma independent Analyst and publishers(**Performance of Drug in Market**)
7. Guidelines and rules set by Government and regulation agency

Identifying Static Sources

Static Data sources :

1. Distributors and their Performance
2. Sales Representatives and Performance
3. Drugs sold out – Day-to-Day Sales
4. Corporates/companies buying new drugs
5. Other Market competitors with similar drugs
6. Clinical Trials
7. Pricing of the Medicines
8. Patient Testimonials
9. Sales Operations and CRM Software
   - Collate data about customer relationships.
   - Documentation of sales team performance.
   - Record customer’s decision-making process.
   - operational costs for sales operations and CRM

Identifying Dynamic Sources

Dynamic Data sources :

1. Social Media – (like : Twitter, Quora)

2. Collating Data from Pharma independent Analyst and publishers(**Performance of Drug in Market**)

3. Guidelines and rules set by Government and Non - Governmental Pharma regulative agency

4. Country/Regional Demographic information

Data Model Selection

Data Model is totally dependent on the type of the incoming data and the way ProPharma is willing to use it for further Analysis.

The data model considered is Hybrid , as the single Data model will not suffice this requirement due to complexity of data around in building this operational pipeline.

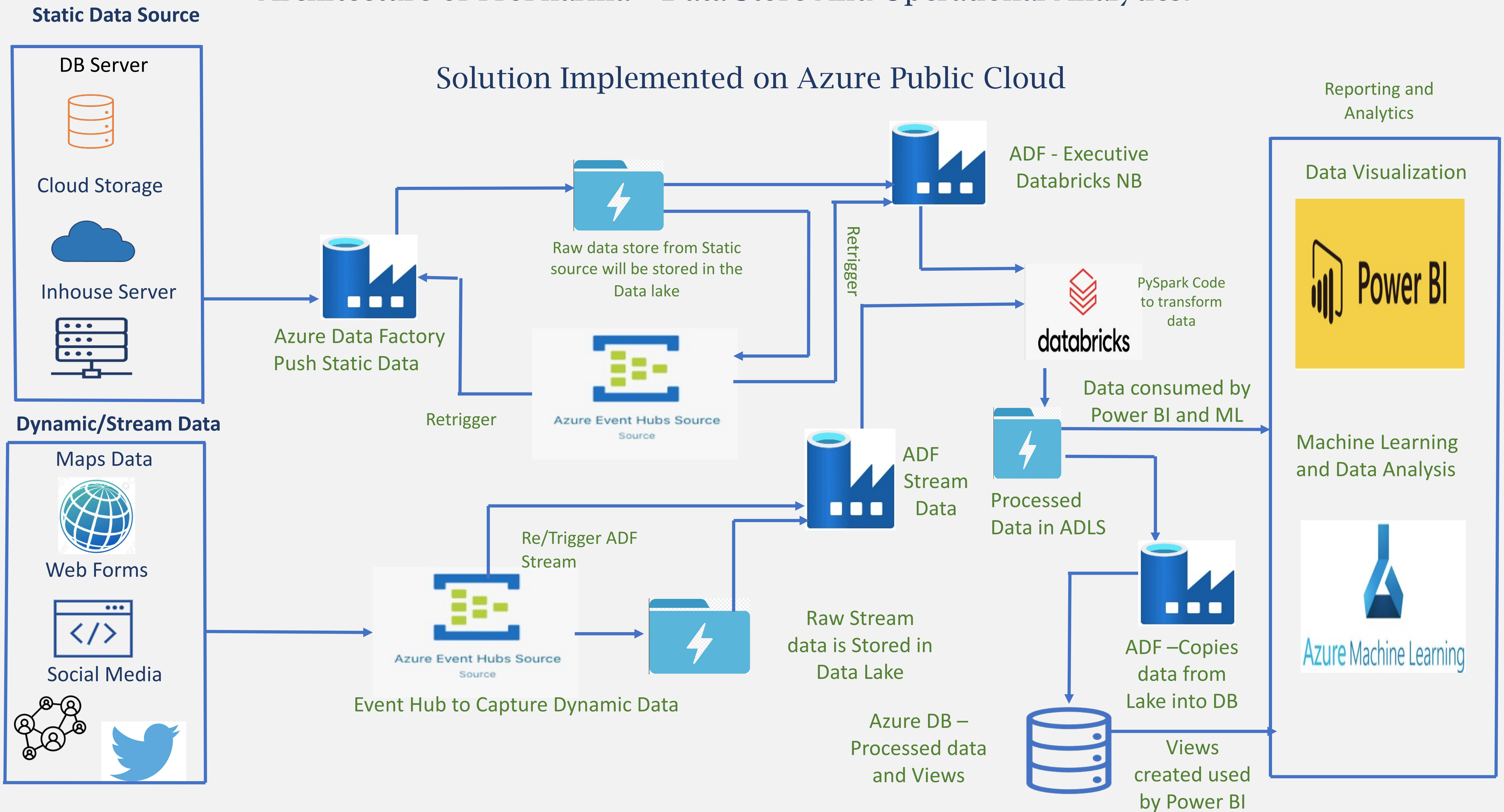
Data sources considered are both static (batch data) and Dynamic(Stream Data)

We have considered a combinational data model that can accommodate the following Models

- Relational Data Model,
- Hierarchical branch Model &
- Network/Graph Model

Architecture of ProPharma – Data Store And Operational Analytics.

Solution Implemented on Azure Public Cloud



Static Data Source

DB Server

Cloud Storage

Inhouse Server

Dynamic/Stream Data

Maps Data

Web Forms

Social Media

Azure Data Factory
Push Static Data

Retrigger

Raw data store from Static
source will be stored in the
Data lake

Retrigger

ADF - Executive
Databricks NB

PySpark Code
to transform
data

Data consumed by
Power BI and ML

Processed
Data in ADLS

ADF
Stream
Data

Re/Trigger ADF
Stream

Event Hub to Capture Dynamic Data

Raw Stream
data is Stored in
Data Lake

ADF –Copies
data from
Lake into DB

Azure DB –
Processed data
and Views

Views
created used
by Power BI

Reporting and
Analytics

Data Visualization

Machine Learning
and Data Analysis

ELT Process and Data load

- Due to multiple sources, which are both static and dynamic. The pipelines that load the Static and Stream data, should be developed individually.

- Pipelines that are going to load **Static Dataset**(changing daily/infrequently), can be loaded through a batch process in Data Lake in batch wise manner.

- For **Dynamic data loads(Streamed data)** that need to be on high alert like **Social media data and Adverse events** about the new drug should be monitored continuously monitored.

- Before the **Stream/Dynamic data** is consumed by the **ELTL** jobs, we need to pass it to Data Lake/Storage layer to capture the data.

- This could be achieved using the Cloud service like **Event Hubs, IOT hubs, AWS kinesis** or pub/sub services

- Once this **Stream/Dynamic data** is stored into RAW Data lake , we can use it as part of ETL pipeline to load and transform data.

- For Data refresh and pipeline failures, since we will employ two separate ETL workflows, will load data, based on type of source dataset and its' functionality, into respective target location into RAW Azure Data Lake system, any failure on the particular pipeline can be easily identified, we can use full load and Incremental load accordingly.

ADF Pipeline Runs & Schedules

❖ Data once loaded in **RAW form into Azure Data lake**, can we be consumed anytime independent of Static/Dynamic data source.

❖ For **Static Data Source**, we can **schedule and configure** linked services in ADF accordingly to **Extract and Load** data from static sources and load that data into **STATIC RAW ADLS FOLDER.**

❖ For the **Static Dataset mentioned**, we will be performing a **Full load in batch process** into Data Lake.

❖ For **Dynamic Data source like social media or Web forms**, once the Tweet/Event happens , Event Hub triggers ADF pipeline to pull data from RAW ADLS Folder.

❖ **Azure Data Factory pipeline** created to handle the Stream will **trigger a Databricks Notebook**(PySpark) to load raw data in ADLS RAW folder into ALDS Processed folder

❖ The Azure Data factory can be configured to execute Databricks PySpark code to perform Relevant Transformation on the incoming data from **STATIC and STREAM RAW** ADLS Folders

❖ The Azure Data bricks, first **processes** the data and ensure the data is loaded into **Processed Azure Data Lake resource**

❖ Finally, we are using one more **Azure execute pipeline**, which will For-Each and copy data activity will iteratively copy files from Processed **ADLS into Azure SQL** Server Data Base.

Type of Load on Static and Dynamic Data Sets

- On the Static incoming data sets , the system will perform a full batch load and push the data into processed Data lake through Databricks and finally into Azure SQL.

- Incase of Failure in Static Pipeline, we can place the Event hub monitoring for the required log file in RAW ADLS and then retriggers incase of pipeline Failure.

- On the Dynamic Data sets coming in, the event hub waits for the event and will load the data into ADLS and in parallel also , send triggers ADF to load data for processing.

- This Stream data incoming will be captured and Hashed to get the uniqueness of Data. For instance, two twitter tweets having redundant information/compliant – will be marked and compared using Hash value and stored in the ADLS processed layer.

- This way we can group the reviews/performance/Market growth of the Drug at granular form to provide meaningful insight.

- Incase of Failure in Stream/Dynamic Pipeline, we can place the Event hub, which keep looking for the required log/touch file in RAW ADLS and then retriggers incase of pipeline Failure.

Reporting and Analytics

- As Azure provide Reporting Tool Power BI as part of cloud service, We are leveraging the same tool for Dashboard creation and Operational Reporting for Business users.

- Power BI will pull data from the Database through views created on Database tables and will not have load the system with Extensive Reads

- Reports are scheduled in Power BI admin console to refresh based on the latest data At scheduled interval of time

- There is also, for further Analysis , for which Azure ML services can be engaged to perform Machine Learning on the incoming data , Refine and Processed data in the SQL to get more meaningful insight.

- ML Models are configurable to run at Sleep hours, so that ETL or Reporting is not affected with this process.

Annexures

https://docs.microsoft.com/en-us/azure/data-factory/

https://docs.microsoft.com/en-us/azure/databricks/

https://docs.microsoft.com/en-us/azure/event-hubs/

https://docs.microsoft.com/en-us/azure/?product=popular

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction

https://docs.microsoft.com/en-us/power-bi/