Case Study

Data Engineering
with PySpark and Azure Services

Case Study

Business Requirements

- We are given a large dataset from a telecom services provider which contains the following data.

  1. **Call Record Details:** A dataset capturing call logs, including caller and receiver details, call duration, and timestamps.
  2. **Customer Usage Data:** This dataset sheds light on individual customer data usage, encompassing voice minutes, text messages, and data consumption.
  3. **Billing and Payment History:** A dataset chronicling billing information, payment dates, and outstanding balances for each customer.
  4. **Network Performance Metrics:** A collection of data spotlighting network performance aspects, encompassing signal strength, call drop rates, and data transfer speeds.
  5. **Customer Complaints:** A dataset meticulously recording customer complaints, categorizing them, detailing resolution times, and assessing customer satisfaction.

Case Study

Business Requirements

- The data set consisting of these details will be available in delimited format and/or semi-structured format in Azure blob storage.
- We need to clean up the files typically by:
  a) Checking for null values and dropping the records in these cases.
  b) Doing data consistency checks such as negative values for instance in fields like data usage or call minutes in the case of Customer Usage Data file.
  c) Also generate a unified view for billing reports
- These tasks are to be performed with Azure Data Factory and PySpark data-frames using data frame function or SQL or a combination of both
- Then the curated data needs to be stored in as the Gold standard for final consumption for reporting and visualization.

Case Study

Data Dictionary

- Data Dictionary for each input file with sample records is as follows.
  - **Customer Usage Data:**

CustomerID	Date	DataUsage	VoiceMinutes	TextMessages
48807980-8f98-4b82-ae94-4ee2d53d6fbc	2023-09-20	1592	121	51
74c040f3-44a4-4878-bea7-fe19233e2073	2023-09-20	3679	302	107

  - **Call Records:**

CustomerID	Caller	Receiver	Date	Duration
6cfed869-965a-4455-a139-e09429673a18	(03) 2225 7374	687-6737	9/3/2023 12:58	74
b0d93e7f-43a8-47d4-aa18-180bdeebd645	(03) 3365 4337	314-3185	9/8/2023 19:41	44

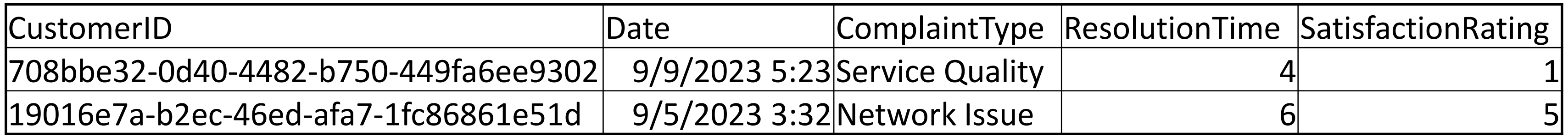
  - **Billing History:**

CustomerID	BillingAmount	PaymentDate	OutstandingBalance
48807980-8f98-4b82-ae94-4ee2d53d6fbc	174.05	2023-08-10	49
74c040f3-44a4-4878-bea7-fe19233e2073	314.05	2023-08-10	35

Case Study

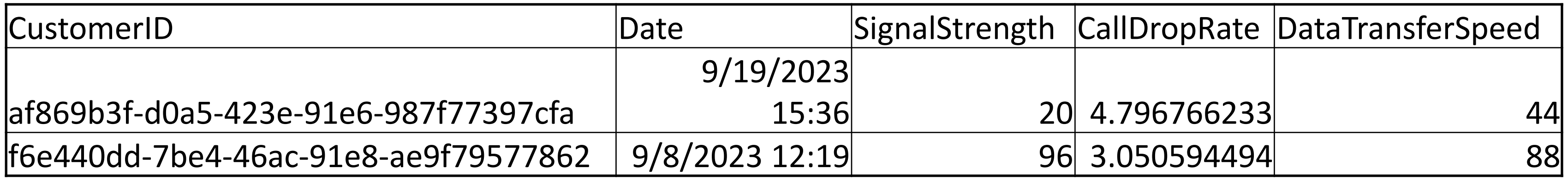
Data Dictionary

- **Complaints Data:**



CustomerID	Date	ComplaintType	ResolutionTime	SatisfactionRating
708bbe32-0d40-4482-b750-449fa6ee9302	9/9/2023 5:23	Service Quality	4	1
19016e7a-b2ec-46ed-afa7-1fc86861e51d	9/5/2023 3:32	Network Issue	6	5

- **Complaints Data:**



CustomerID	Date	SignalStrength	CallDropRate	DataTransferSpeed
af869b3f-d0a5-423e-91e6-987f77397cfa	9/19/2023 15:36	20	4.796766233	44
f6e440dd-7be4-46ac-91e8-ae9f79577862	9/8/2023 12:19	96	3.050594494	88

Case Study

Problem Statement

- The data needs to be cleaned and pre-processing tasks need to be done.
- The following reports and visualizations (dashboard) needs be generated from the curated data.
  - Monthly Billing report for current month: This is required to be generated by extracting the details from the raw data set – Customer Usage and Caller Records.
  - From Network Metrics and Caller Complaints data, after initial clean up as required, visualizations needs to be generated as dashboards containing bar charts and other graphs as appropriate for the metrics being presented.

Case Study

Problem Statement Details:

- Billing history is available from January till August.
- Monthly billing report for September is required to be generated by extracting the details from the raw data as below.
  - The call records data needs to be cleaned up by removing invalid records in which the duration zero or negative.
  - The duration of calls for each customer has to be added to the voice minutes of customer usage data.
  - Payment amount is to be calculated using the formula: DataUsage*0.1 + VoiceMinutes*0.01 + TextMessages*0.1
  - From the August month's billing history the outstanding balance of each customer needs to be added to the above calculated value to get the bill amount for September.

Case Study

Problem Statement

- Complaints data has several records with empty columns i.e. nulls
- These need to be filtered out before the following visualizations are generated.
  - Bar charts for each type of complaints showing average resolution time and satisfaction rating
  - Gauges showing the over all min, max and average resolution time and satisfaction rating
- In network metrics data wherever the metrics are not available the columns are filled up with NA.
- These records which need to be removed and the following visualizations are generated.
  - Gauges showing the min, max and average of each the metrics

Case Study

Solution Approach:

- Based on the problem statement it may be clear that most of the requirements are similar to the classroom exercises done on PySpark.
- For the clean up and other pre-processing transformations you need to identify where ADF can be used and where PySpark has to be used.
- The raw data will be made available as CSV files in Azure blob storage.
- You need to use medallion architecture for the raw, processed/curated and business ready datasets prepared in the above processes.
- The data in Silver and Gold zones can be stored in Parquet format.
- Power BI can be used to generate dashboards from the above data.

Case Study

Solution Approach:

- The case study solution should contain:
  - Your understanding of the problem statement and requirements
  - Design or architecture diagram of the solution specifying the storage repositories of the source data, the curated data and final reports
  - PySpark application code and scripts of any other tools if used
  - The class will be divided into 4 teams and within the teams the tasks can be distributed. However each team member should have the complete picture and understanding of the problem and the solution.