AIMIA Data Engineer - Technical Challenge

# **Data Pipeline**

# **Data Processing and Data Science**

Upload the \*.gz files to HDFS/S3 and store as read only raw data.

Use Spark Databricks CSV package to read the \*.gz files

1. Data Processing: Use Pivot

The input data is narrow table i.e. we must pivot the data in order to calculate the statistics

Use Spark **dataframe Pivot function**

Extract only certain features that seem relevant:

“STATION”, “DATE”, "TMAX", "TMIN", "TOBS", "PRCP", "SNOW", "SNWD"

Create static views of data i.e. required statistics are computed and stored to minimise the number of iterations through the data

**NOTE: Sample code is attached. TESTED using SPARK Scala REPL**

1. Data Visualization

Trends in the data can be visualised using tools such as ggplot2 in R.  
Visualisation will help in analysing the patterns and relationships between different data variables.

1. Data Science Application
2. The transformed data can be aggregated to produce more stats. For example, find median temperature for all years for each station, region, country etc.
3. Example use cases:
   1. Find countries,regions, states with similar weather patterns

As the number of variables are less we can use euclidean distance

* 1. Predict weather behaviour

Note: Used R to show sample scripts