**Tools Used:**

AWS S3, EMR with spark( scala ), Redshift

**Why EMR ?**

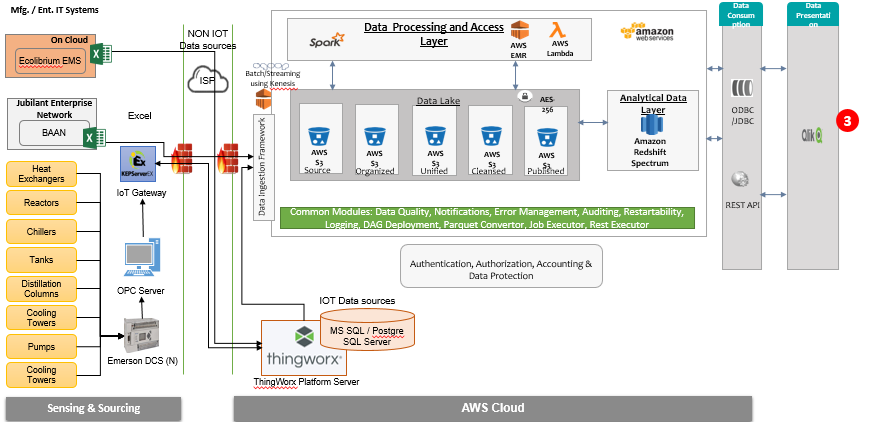
Amazon EMR is based on Apache Hadoop, a Java-based programming framework that supports the processing of large data sets in a distributed computing environment.

Amazon EMR processes big data across a Hadoop cluster of virtual servers on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3).

The elastic in EMR's name refers to its dynamic resizing ability, which allows it to ramp up or reduce resource use depending on the demand at any given time.

Amazon EMR is used for data analysis in log analysis, web indexing, data warehousing, machine learning, financial analysis, scientific simulation, bioinformatics and more.

**Technical overview:**



1. Customers will send SAP Reports on daily basis via BAAN system and it will be stored in S3 buckets.
2. By running pyspark jobs in EMR cluster Data transformations like Excel to CSV conversion will happen.
3. Data processing spark jobs will be submitted in EMR cluster for different usecase
4. Processed data will be stored again in different S3 buckets.
5. By running pyspark jobs in EMR cluster data copy will happen from s3 into Redshift Database.
6. Qlik dashboards connects to the Redshift warehouse.

**Use case:**

**Use case 1: Inventory KPI Visualization dashboard**

                     Day to day inventory closing balance & inventory value (in cost) calculation for the goods( raw material, semi-finished and finished)

**Use case 2: Production performance KPI Visualization dashboard**

                    Analyzing the variance of actual Vs Planned production of the finished goods

**Use case 3: Energy management dashboard**

                     Calculating the specific energy consumption and variance for the produced goods

**Process Flow:**

Input Data -> S3 Storage -> Data processing using EMR with spark -> Data store in Redshift -> Output to Qlikview

Stored input data received from client in s3 buckets and processed the data using EMR with spark . Finally pushed the data into AWS Redshift for visualization.

1) Received files (.xls) from Jubilant

2) We convert them into .csv programmatically using pyspark and store them into S3 raw layer

3) Transformations done using Spark/Scala in EMR and the processed data in the S3 publish layer

4) Copy data from publish layer to Redshift

5) Qlik dashboards connects to the Redshift warehouse.

**Lessons Learned:**

Learnt how to overcome performance issue while ingesting processed data into redshift database.

Learnt how to create EMR with spark cluster dynamically based on use case requirement.