Context:

A Pizza Restaurant chain “Pizza House” has more than 2000 stores across the country. Each store manages its own inventory of raw materials. Each store prepares pizzas, side dishes, etc. and sells them along with ready to eat products such as cookies, drinks, etc. The sale can happen by Point of Sale (POS) or Online. The online transactions would be flowing in real time whereas the transactions made by POS can be synced every 15 minutes in batches. They offer pick-up and deliveries by 3rd party providers.

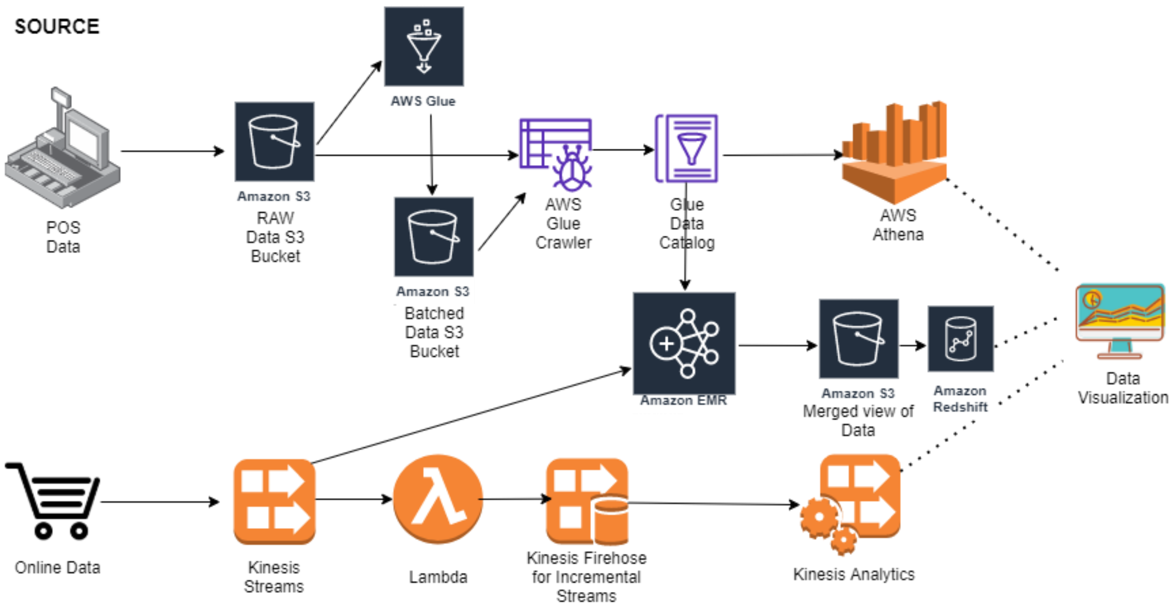
At the head office of the restaurant chain, management is concerned with the logistics of ordering, stocking and selling products while maximizing profits as well as understanding their marketing & communications. Several promotional schemes such as temporary price reductions, ads in newspapers, displays etc., also keep rising. Considering the huge data volumes (hundreds of GB per month) and the variety of the data they have; management wants the architecture to be robust enough to handle the varying data loads.

Design a cloud data platform to process and deliver insights based on the above. Please provide a high level solution design for the architecture. Feel free to choose any cloud provider you want.

Requirements

1. Handle large write volume: Billions of write events per day.
2. Handle large read/query volume: Millions of merchants wish to gain insight into their business. Read/Query patterns are time-series related metrics.
3. Provide metrics to customers with at most one hour delay.
4. Run with minimum downtime.
5. Have the ability to reprocess historical data in case of bugs in the processing logic.

High Level Solution Design:



Explanation for POS Data:

POS Data is being sent to S3 bucket which is then processed using AWS Glue Job.

AWS Glue Job sends the POS data to EMR for merging both sources of Data.

AWS Athena is being used for analysing the data for visualization.

Explanation for Online Data:

Online Data Streams are being processed through Kinesis Streams.

Also, the Kinesis stream data is being consumed by Amazon EMR for merging both the sources of Data.

Lambda is used to sending data to Kinesis Firehose for Incremental processing of Streams.

Kinesis Analytics is being used for analysing this data .

Merged Data:

A data ware house has been created in Amazon Redshift to store both data at one place so that reporting and other analytics can be done.

Explanation of the requirements:

1. Handle large write volume: Billions of write events per day.

   Above design would efficiently handle billions of write events per day as we are separating Online vs POS data processing mechanism and choosing the architectural objects which have been made to cater such requirements.

2. Handle large read/query volume: Millions of merchants wish to gain insight into their business. Read/Query patterns are time-series related metrics.

   As read has been separated from write , there is no bottleneck in the system which would cause issues with parallel read and writes.

3. Provide metrics to customers with at most one hour delay.

   This system should be able to provide metrics with max 15 mins delay.

4. Run with minimum downtime.

   No downtime as such.

5. Have the ability to reprocess historical data in case of bugs in the processing logic.

   As we are loading the data incrementally , the cloudwatch logs and also Kinesis,Lambda would take care of not loading the same batch again and hence would suffice the requirement.

Table Structure:

In the Datawarehouse, we will create the denormalized tables as per below:

Dimension Tables:

DimEmployees

DimCustomers

DimSuppliers

DimProducts

DimDate

Fact Tables:

FactSales

FactOrders

FactPayments