SkipTheDishes Tech Challenge

Data Engineer

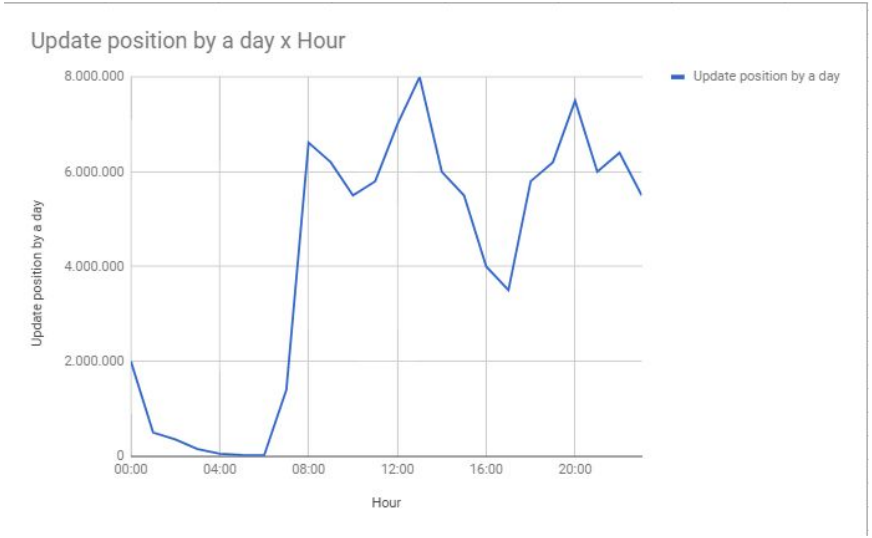
Miller Carvalho

Topics

- Database Solution
- Identify areas to expand the service
- Moving environment to the Cloud

Database Solution

● Sample Graph Estimation - Update position by a day x Hour graph:



Database Solution

- Use DynamoDB - multi-model, schemaless, scalable, auto-managed and provisioned read/write capability @ AWS;
- Table Model:

Table: CouriersPosition

CourierId:Number *(Partition Key)*	PositionTimeStamp:String *(Sort Key)*	Area
1	2018-03-17T12:22:34Z	Eastman
2	2018-03-17T12:25:55Z	Eastman
5	2018-03-17T12:26:13Z	Parkland

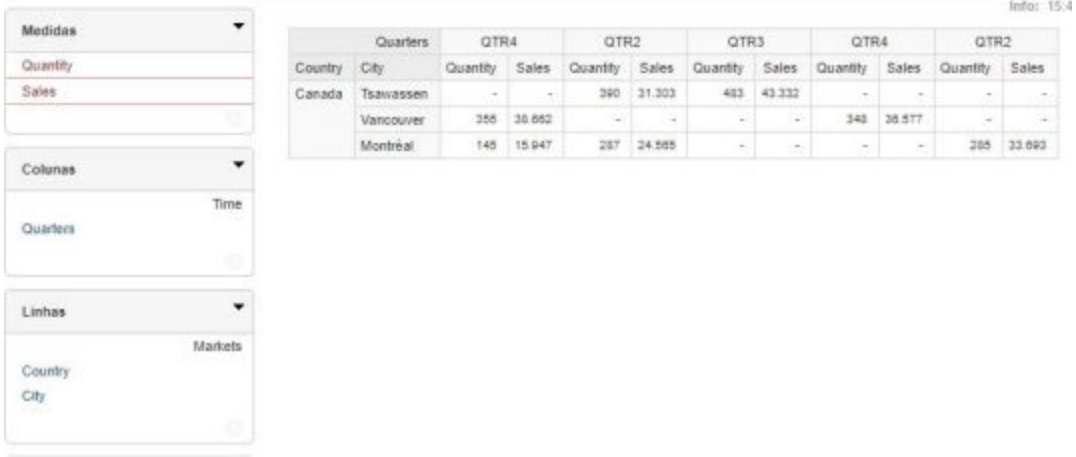
Database Solution

Considering that will be exponential growth we could:

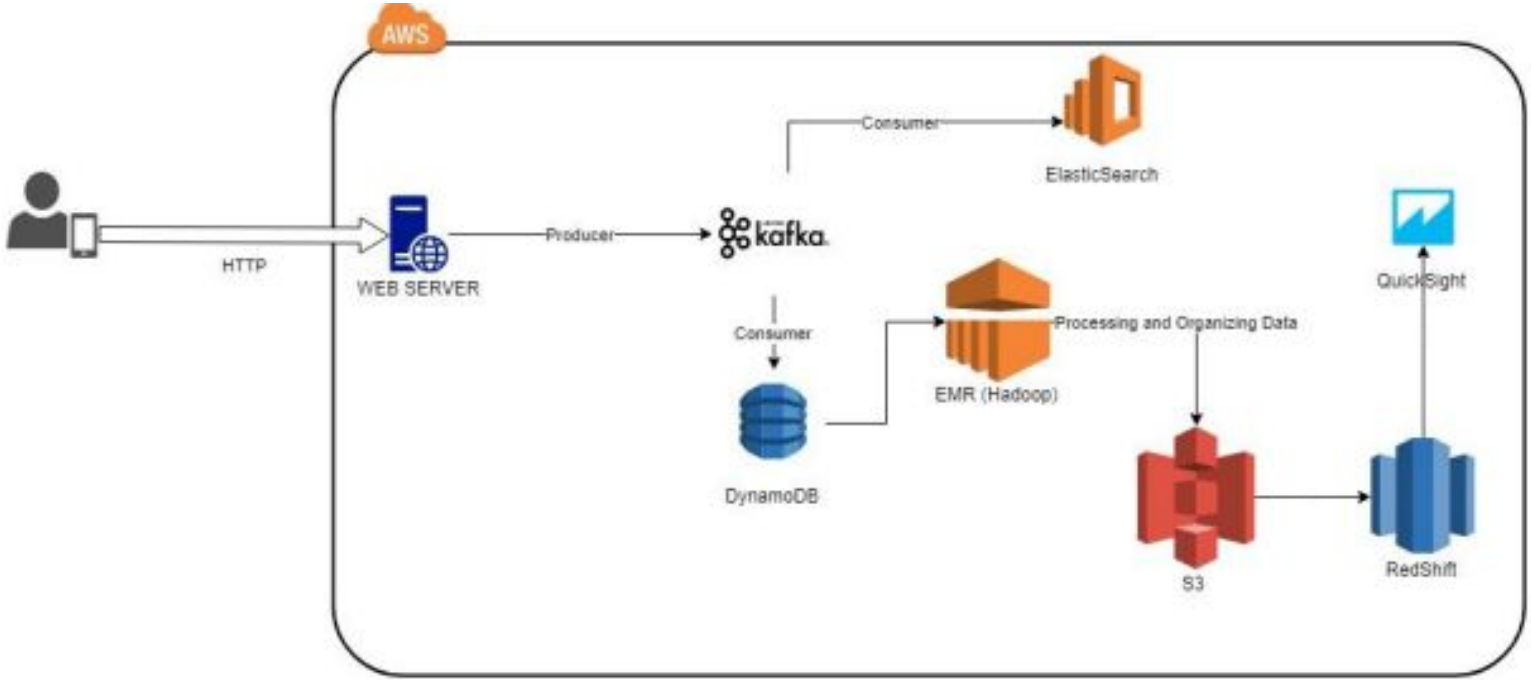
- First: auto-scaling - could be expensive, but you can enable just to see the right behavior;
- Second: DynamoDB must be able to write 3.333 events per second.
- It will works if you provision 3.333 units to write per second in a table that will modeled to able cover courier’s system for the next 6 months.

Identify areas to expand the service

- You can use Redshift to copy DynamoDB's data, and you could perform SQL queries;
- You can use a OLAP engine like Pentaho Mondrian to see your data in multidimensional model (cubes), like image below:



Moving environment to the Cloud



Moving environment to the Cloud

- Kafka to stream events. You should keep the last 24 hours of events for a lot of consumers that wants to make near real-time decision or insert data into relational databases or DynamoDB;
- DynamoDB to persist data;
- Hadoop (EMR) to process distributed events in large scale, organizing data and record them into S3;
- Redshift (Columnar Database) will get S3's organized data to datawarehouse purposes (S3DistCP);
- QuickSight for analytics, reports and dashboards;
- Elasticsearch to keep events to be able to full text search events (for segmentation issues).