Project : Spark Streaming Application

Description : You are working as a data engineer for one major e-commerce client which has huge infrastructure setup done at their end.

Whenever any production change happens, it is first raised in CAB meeting (Change approval Board) meeting, discussed and approved and then only the change goes to production.

Change may be

* Change in os version i.e. os upgrade
* New DB installation on server
* New Middleware installation
* New Network component installed

A master information is maintained to keep a track of all information associated with all the production servers.

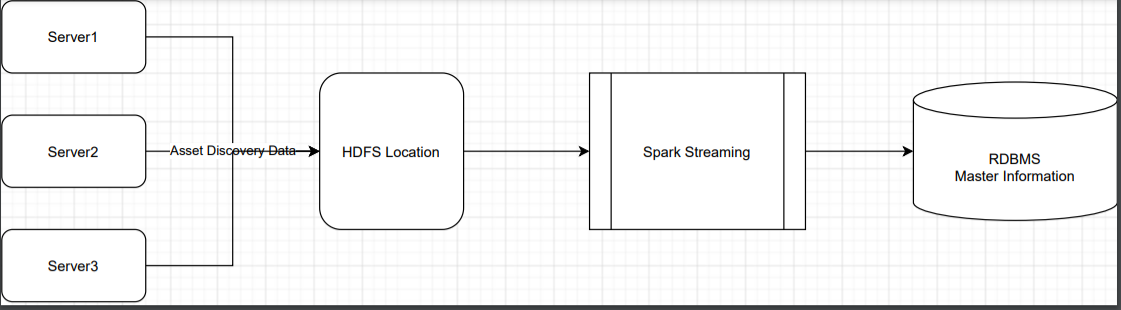
Client is having more than 40K servers.

Problem statement : Even though all the changes are approved in CAB, there are some deviations i.e. production changes are happening without CAB as well.

To control and keep a track of these ad hoc changes, new project was launched for asset discovery.

We get scripts running on all production servers to track what all is running on different servers of production.

This is a real time monitoring/asset discovery which keeps on happening and it is like as follows.



Action : Produce the files having following content (sample record)

Assetid | 12345, TechStack | OS:RHEL;Version:6.8.7;DB:MongoDB;Middleware:Nginx,additional\_info | IP:10.20.30.44;Hname:hname1;owner:abc@xyz.com

Assetid | 12346, TechStack | OS:CentOS;Version:3.8.2;DB:Cassandra;Application:XYZApp,additional\_info | IP:10.20.30.45;Hname:hname2;owner:pqr@xyz.com

Spark streaming application will capture the data on a streaming window of 15 secs. All the files produced on real time basis should be captured in micro batch processing manner and need to produce a dataframe with following column.

AssetID, Header, Key, Value

Show the dataframe in console.

Sample output :

AssetID, Header, Key, Value

12345, TechStack, OS, RHEL

12345, TechStack, Version,6.8.7

12345, additional\_info, IP, 10.20.30.44

12346, TechStack, OS, CentOS

12346, TechStack, Version,3.8.2

12346, additional\_info, IP, 10.20.30.45

12346, TechStack, Application,XYZApp

12346, TechStack, DB,Cassandra

12346, additional\_info, Hname, hname2

12346, additional\_info, owner, pqr@xyz.com

High Level Design

