Recent project description

My recent assignment was with Walmart,Bentonville, Arkansas. I was a data engineer working in Directed diagnostics which is part of Global data analytics, project is about predicting the purchase orders (no po, future po, late pos) in advance and displaying the lost sales for itemlevel to stake holders(buyer, replenishment managers,demand managers)

Data pipeline includes(Mongodb, kafka, hadoop,hive,Mariadb(similar to sql server), UI) like we get data from legacy systems like Teradata , mainframes to Mongodb and from mongodb data via kafka queue data is landed onto data lake (hadoop). Records are in semi structed format which are json records and the data is converted to table in Hive . On hive according to business requirement , I have developed hive queries and exported the data onto Mariadb(consumption layer). To export/import the data walmart has their own internal tool called Aorta framework which works as Sqoop . Aorta framework file format is yaml, yaml have its own set of rules defined like source, target , query . **Invoke hive queries, Invoke sqooping data, Invoking Linux shell scripts** is done with aorta framework. We have used Automic to schedule jobs, it’s a daily load of data . Aorta yamls that have hive queries and sqoop queries and these are scheduled on Automic scheduler. Worked on Thoughtspot a business analytics tool to do highlevel computations(in place of mariadb). We have 2 weeks sprint and we used to deliver deliverables for every 2 weeks to business. Actively participated in sprint demos, sprint planning and sprint retrospectives.

Coming to spark, in my previous project at stealth mode fintech startup worked with spark.We have tera bytes of data coming from various sources like mysql, oracle ,teradata. I have used sqoop to ingest the data on to Hadoop , on hadoop using spark , I have performed computations using **transformations**(eg:map,filter,join) and **actions**(count,collect,foreach(println),top,take),Used spark with scala, converted json files ,csv files, text files to data frames and applied logic upon data frames .

Opted spark upon hive because of processing time, spark took 30mins to process 1 terabyte of data where as hive taken 24 hrs to process 1 Tb of data.

Personal details:

Shailaja koppishetty

Btech jntu, Ms software engineering Villanova university

Current location:Bentonville, Arkansas(cst)

Have 6+ years of experience as data engineer.