Describe your experience?

Hi, my name is Dave. I am currently working for wells fargo in san franciscon, california as a data scientist.

I have been in the field of data science for more than 5 years since being working for Axiom Tech Group,

The Predictive Analytics Project I'm working on is supposed to to improve the prediction of misconduct through the use of Machine Learning (ML) and Natural Language Processing (NLP) techniques. We gain insights gleaned from data that were used to determine target markets and their issues, plan supply and demand of financial products and ways to save on costs with predictive analytics as well as reduce risk and examine strategic partnerships.

I am very familiar with data pipeline building and deployment of different machine learning and neural network using a variety of systems, these parts are what i am in charge of. My industry experience includes predictive analytics in consulting, marketing,

finance, logistic area.

My project is wrapping up, about to hand over to devs op team. we made good progress. so i'm seeking for a new opportunity. I hope my solid background and experience can contribute to your success.

What the Pipeline at the company you work at for DS?

We usually use google cloud platform Data Pipeline, it is a web service that you can use to automate the movement and transformation of data. With AWS Data Pipeline, you can define data-driven workflows, so that tasks can be dependent on the successful completion of previous tasks. we define the parameters of data transformations and gcp Data Pipeline enforces the logic that we’ve set up.

Your Role at the company? Did you help design the pipeline

I’m a data scientist, I was involved in building the pipeline. We used spark to build the pipeline.

Why would Spark be used in this pipeline?

Spark ML Pipelines provide a uniform set of high-level APIs built on top of data frame that help users create and tune practical machine learning pipelines.

What is a schema?

A schema is the description of the structure of your data (which together create a [Dataset](https://jaceklaskowski.gitbooks.io/mastering-spark-sql/spark-sql-Dataset.html) in Spark SQL). In structural data, It tells us information of each column, like datatype, what it is about.

What file format did you use for your pipelines?

We mainly deal with json file. Some times .csv file.

Rate yourself spark on spark 1-5?

I’ll give 4.

What is an RDD?

RDD is a collection of elements partitioned across the nodes of the cluster that can be operated on in parallel. RDDs are created by starting with a file in the Hadoop file system (or any other Hadoop-supported file system), or an existing Scala collection in the driver program, and transforming it.

What is an executor in spark?

Executors are worker nodes' processes in charge of running individual tasks in a given Spark job. They are launched at the beginning of a Spark application and typically run for the entire lifetime of an application. Once they have run the task they send the results to the driver. They also provide in-memory storage for RDDs that are cached by user programs through Block Manager.

How do you do Tuning of spark?

Tuning Resource Allocation

## Tuning Parallelism

## Slimming Down Your Data Structures

## Data Formats

What are Narrow Dependencies?

Narrow Dependency is a base (abstract) [Dependency](https://jaceklaskowski.gitbooks.io/mastering-apache-spark/spark-rdd-dependencies.html) with narrow (limited) number of [partitions](https://jaceklaskowski.gitbooks.io/mastering-apache-spark/spark-rdd-Partition.html) of the parent RDD that are required to compute a partition of the child RDD.

What are Wide Transformations?

Wide transformations are the result of groupByKey and reduceByKey. The data required to compute the records in a single partition may reside in many partitions of the parent RDD.

Wide transformations are also called shuffle transformations as they may or may not depend on a shuffle.

|  |
| --- |
|  |

What is a Dag in spark?

Dag is directed acyclic graph, it's a finite directed graph with no direct cycles. It is pipeline. It is an alternative to the mapreduce it is faster since it intermediate results are not written to disk.

What is a Skew Join?

Basically, when there is a table with skew data in the joining column, we use skew join feature. On defining what is skewed table, it is a table that is having values that are present in large numbers in the table compared to other data. However, while the rest of the data is stored in a separate file Skew data is stored in a separate file.