* Single cloud platform where all data needs can be handled i.e., entire data team can collaborate.

Data analyst – try to get the data – can use spark SQL, to integrate visualization tools like tableau, powerbi

-huge amount data- clean, transform and process

DE – migrate our data/create ETL activities to get data, to process data

DB provides faster ETL’s

Supports pyspark, sparkSQL, scala, R

Data scientist – to build models (SparkML), has all libraries

* Built on top of existing cloud platforms like AWS, Microsoft Azure, Google cloud

Data Ingestion – getting data from different sources to one point, then we try to transform, analyze and process data to get some meaningful insight(all this can be done in data bricks). After this it can be sent to ML flow, store it in data warehouse or to some visualization tools.

sudo jps – displays whether demons(java processes/Hadoop & spark resources) are running or not

spark -shell – to start the spark

**CreatIng rdds using Scala**

1. Using parallelized collections

**sc.parallelize(1 to 100, 5).collect()** – spark context paralysed method to create paralysed collection with 1 to 100 in 5 different partitions.