**Adb Day3 Notes**

**Features of Delta Lake**

1. **Acid Transaction**

Although Delta Table is not a Relational Database Table still it introduces the concept of ACID transactions on SPARK

1. **Scalable metadata handling**
2. **Streaming and batch Unification**

Delta table in delta lake can be a batch table as well as Streaming source and Sink

1. **Schema Enforcement**

Delta lake automatically handles the schema variations of data to be inserted into delta tables to prevent insertion of bad records

1. **Time travel**

It has data versioning enabling us for rollbacks

1. **Upsert and Delete**

**Text Input Format**

**Default json,csv**

**Parquet format(Columnar format)**

* Parquet file format stores data in form of columnar file storage
* This means that instead of storing row data adjacent side we can also store column values adjacent to each other
* So, datasets are partitioned both horizontally and vertically.
* It is useful while reading a subset of data stored on disk as it can access all values of a single column very quickly without reading whole records
* Design based on googles Dremel paper
* Schema segregated into footer
* Column major format with stripes
* Simple type model with logical types
* All data pushed to leaves of tree

Avro format

Medical data

Opiniated format that understands data in Hadoop which is not in key/value pairs but in int/string