**Evans Mosomi -100719552**

Q: Processing Service – **DataPrep**

Q: **DIFFERENCES**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Advantages | Disadvantages | Limitations |
| DataPrep | -Relies on automating clusters  -Fouces on UI for clear data processing | -Easy to follow as it follows intuitive design  -Easily convert data form to spreadsheet view  -Handles large amounts of data seamlessly | - Maxed out at 1000 datasets in a database.  -Limited memory size for quick data retrieval  -UTF-32 is unsupported |
| Dataflow | -Relies on automation to create clusters  - Reliable for BigQuery  -Perfect for data ingestion: batch and stream processing | -Quite difficult to use and navigate | - Doesn’t work well with Hadoop clusters |
| DataProc | -Sets up clusters manually  - Great over Hadoop and Apache clusters | -Not suitable for BigQuery | -UI isn’t as seamless for an average user |

Q: LIDAR Dataset for stream and batch processing

|  |  |  |  |
| --- | --- | --- | --- |
| Application | Impact | Used Dataset | Tools |
| LIDAR dataset – creates high-resolution models of ground elevation. It can be used for vulnerability analysis e.g. earthquake zones, farming issues, forestation and land management views. | - Disaster prevention using the vulnerability assessment tool  -Object detection for vehicles  -Motion sensors in smart humans  -Geographical improvements e.g. new properties | Uses a long-term vision dataset in combination with LIDAR i.e. NCLT dataset.  The schemas involved have dates, time, weather patterns, earthquake measurements, rainfall etc. | Machine learning tools such as Apache Spark & Scikit learn  -Hadoop Distributed File System |