Name: Tegveer Singh

ID: 100730432

Google Cloud has another processing service called DataProc. Name another processing service that is usually used in the cloud environment (not necessarily GCP). Compare between it and both Dataflow and DataProc. Your comparison may include but is not limited to the major differences, advantages, disadvantages, and limitations.

Criteria	**DataProc**	**DataFlow**	**AWS Kinesis**
Main Difference	Create a cluster of servers to perform ETL tasks, very similar to how Hadoop would work	The ETL tasks are serverless and use Java/Python/Go Apache Beam SDK to perform Batch and Stream processing tasks	Real-time processing for large amounts of data and runs on AWS EC2 instances
Compatibility	Works by using Spark and Hadoop clusters and can be integrated with other GCP services like BigQuery, BigTable	Works with open source programming framework Apache Beam	Can be integrated with S3, DynamoDB and RedShift
Limits	7500 requests per minute	Max 1000 workers with 15000 messages/30 seconds per worker	Higher transaction limits than GCP services. One call retrieves 10000 records (10MB)
Advantages	- Very low overhead and can handle a large amount of data - Clusters that are not needed can be removed - Integrated with Spark for Machine Learning and Data Science algorithms	- Provides both Stream and Batch Processing - Does not run on top of Hadoop reducing operational overhead - Can Integrate with other services AWS, Hadoop, etc - Can process datasets with unknown sizes due to Autoscaling	- Support large number of consumers - Quick data processing with real-time graphs and analytics - Very high read and write throughput - Kinesis Client Library available for languages like Java

Disadvantages	- Reduction in performance as data increases - High Dependency on Hadoop leading to low portability - No Stream processing - Limited to datasets of known size	- Slightly more expensive in comparison to DataProc - More bound to google technologies	- Higher cost than DataFlow and DataProc as charged per shard per hour - Capability limited by the EC2 instance Kinesis is running on

Suggest a practical application using both stream and batch processing that can be applied to a given dataset. It's expected to use the dataset uploaded in the third milestone but you can use any other dataset. If you decide to use another dataset, It should maintain both variety and huge volume. Your report should include but not limited to

- *The application.*
- *Its impact.*
- *The used dataset (size, schema/structure).*
- *A graph showing the proposed pipeline(s).*
- *List of other tools (AI, clustering,…) needed to implement that application.*

Application: Add labels for vehicles, signs, traffic lights as boundary boxes on the videos

Impact: Autonomous vehicles and urban driving safety

Dataset: The chosen dataset is the CARLA dataset for autonomous driving

Tools: DataFlow, Apache Beam, BigQuery, Tensorflow

Graph: