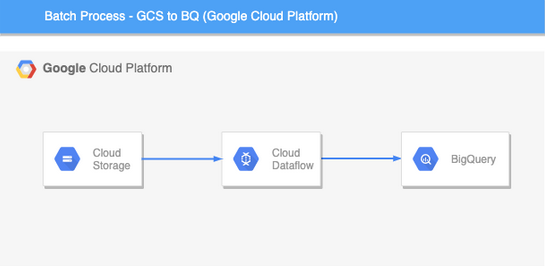
**DataPipeline using Apache Beam and Google Cloud DataFlow as Runner and BigQuery as DataSink**

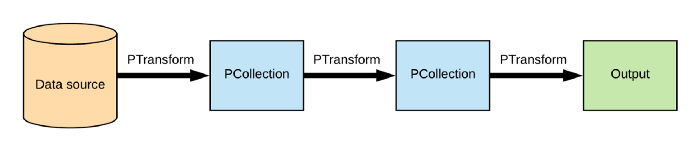


**Goals and Overview**

To Develop a data pipeline to ingest, process, store it so as to access it through different means.

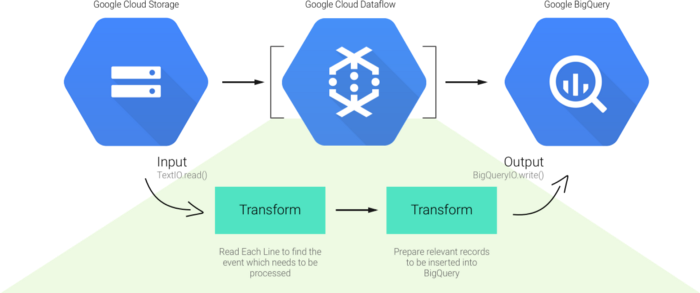
Apache Beam is an open-source, unified model for constructing both batch and streaming data processing pipelines. Beam supports multiple language-specific SDKs for writing pipelines against the Beam Model such as Java, Python, and Go and Runners for executing them on distributed processing backends, including Apache Flink, Apache Spark, Google Cloud Dataflow and Hazelcast Jet.

* **Data pipeline using Google Cloud Platform(GCP) services**



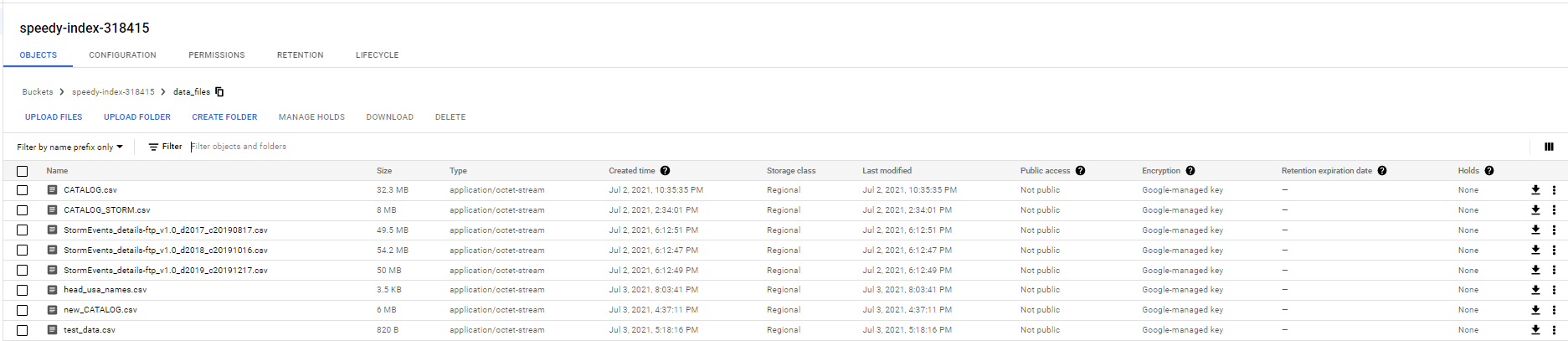
**The Anatomy of a Data Pipeline:**

* **Pipeline:**manages a directed acyclic graph (DAG) of PTransforms and PCollections that is ready for execution.
* **PCollection:**represents a collection of bounded or unbounded data.
* **PTransform:**transforms input PCollections into output PCollections.
* **PipelineRunner:**represents where and how the pipeline should execute.
* **I/O transform:**Beam comes with a number of “IOs” — library PTransforms that read or write data to various external storage systems.
* **Flow of the pipeline**

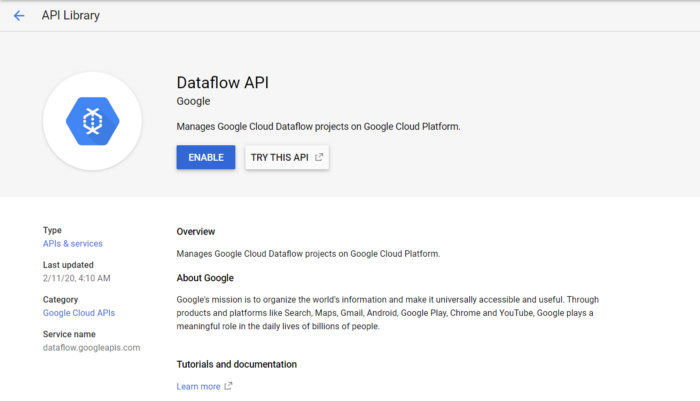


1. Reads the data from google cloud storage bucket (Batch).
2. Apply some transformations such as splitting data by comma separator, dropping unwanted columns, convert data types, etc.
3. Write the data into data Sink (BigQuery)
4. Joins Bigquery tables using reusable Composite Transforms
5. Analyze it using Data studio.

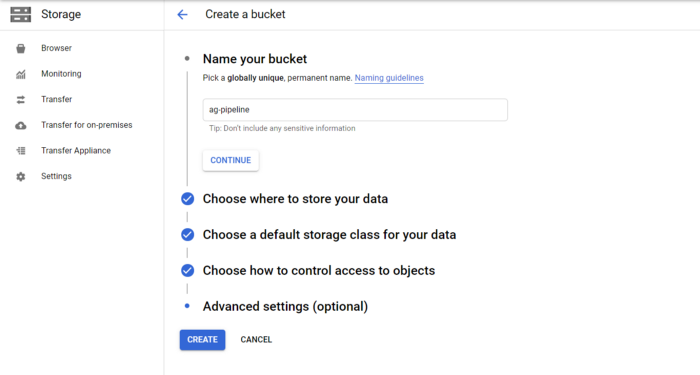
Upload this dataset to google cloud bucket-



Before we run the pipeline, we need to enable Dataflow and Bigquery APIs.



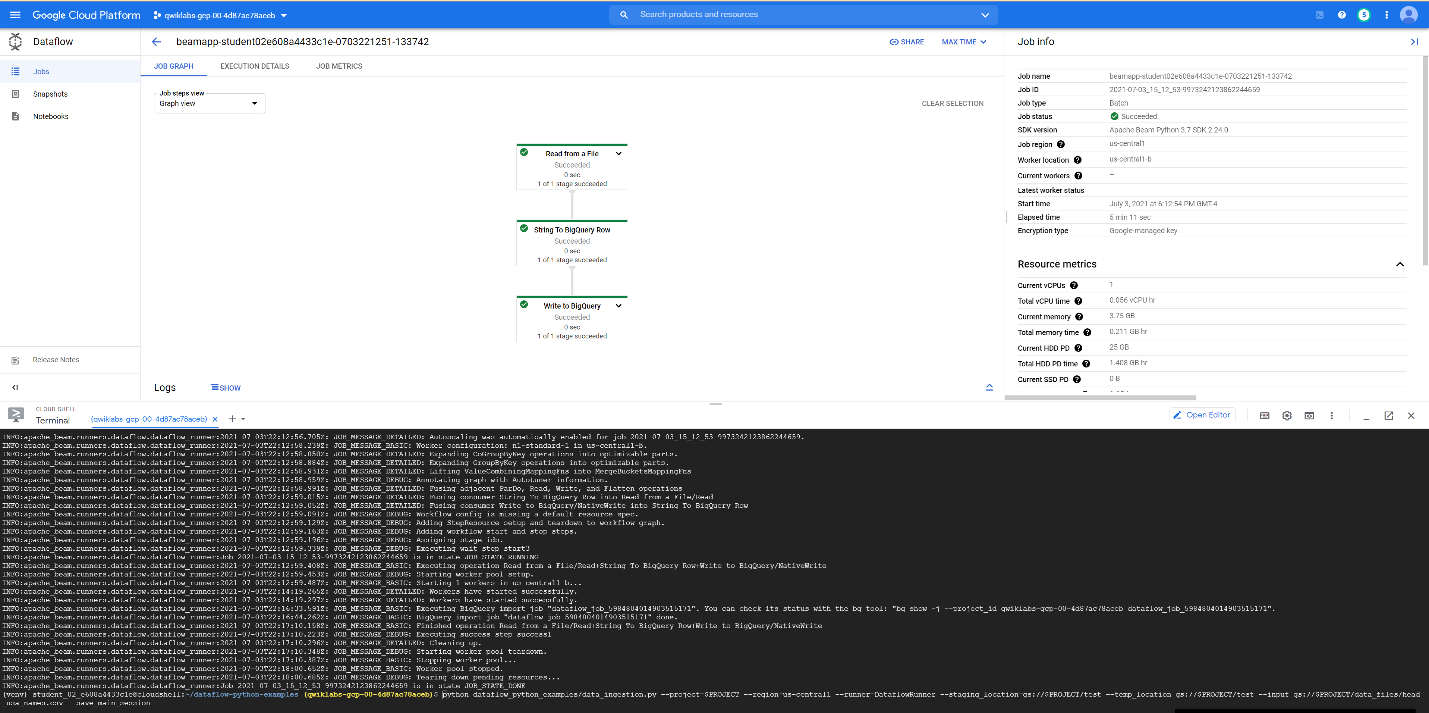
Dataflow will use cloud bucket as a staging location to store temporary files. We will create a cloud storage bucket and choose the nearest location (Region).



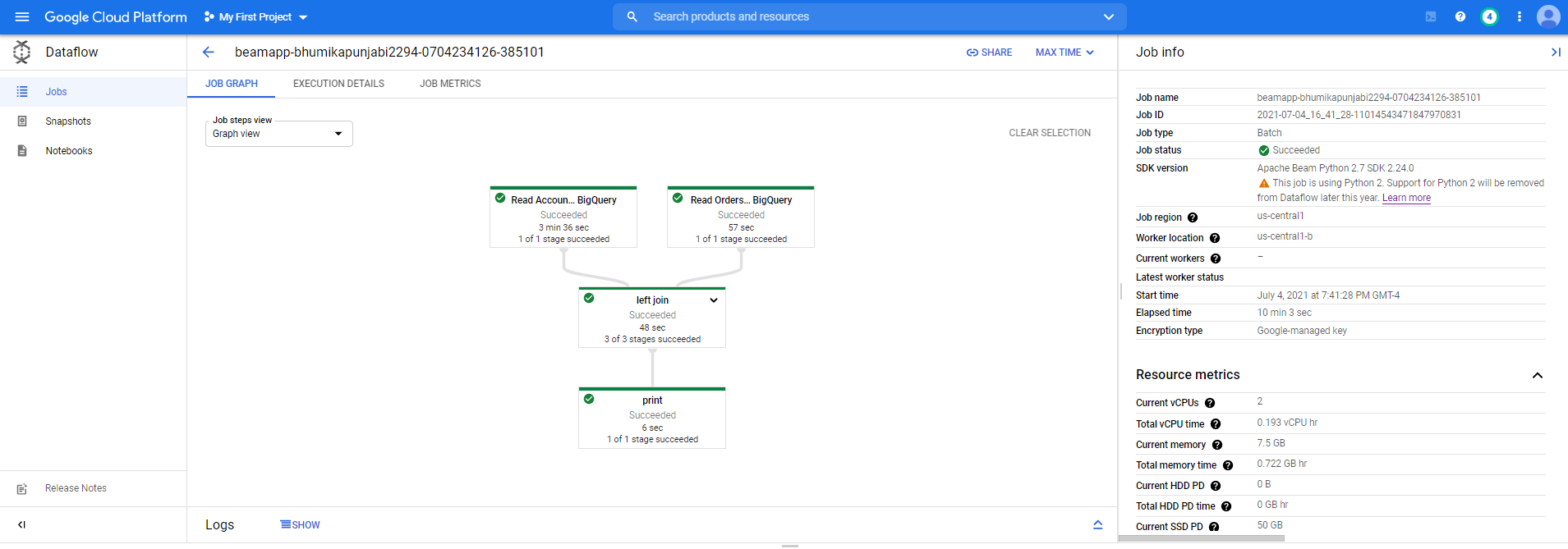
We will create BigQuery dataset and table with the appropriate schema as a data sink where our output from the dataflow job will reside in.

To run the pipeline, we need to have Apache Beam library installed on Virtual Machine.

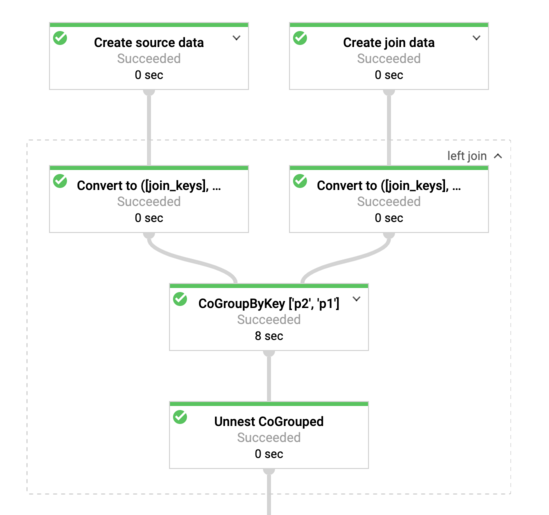
Ingestion Pipeline-



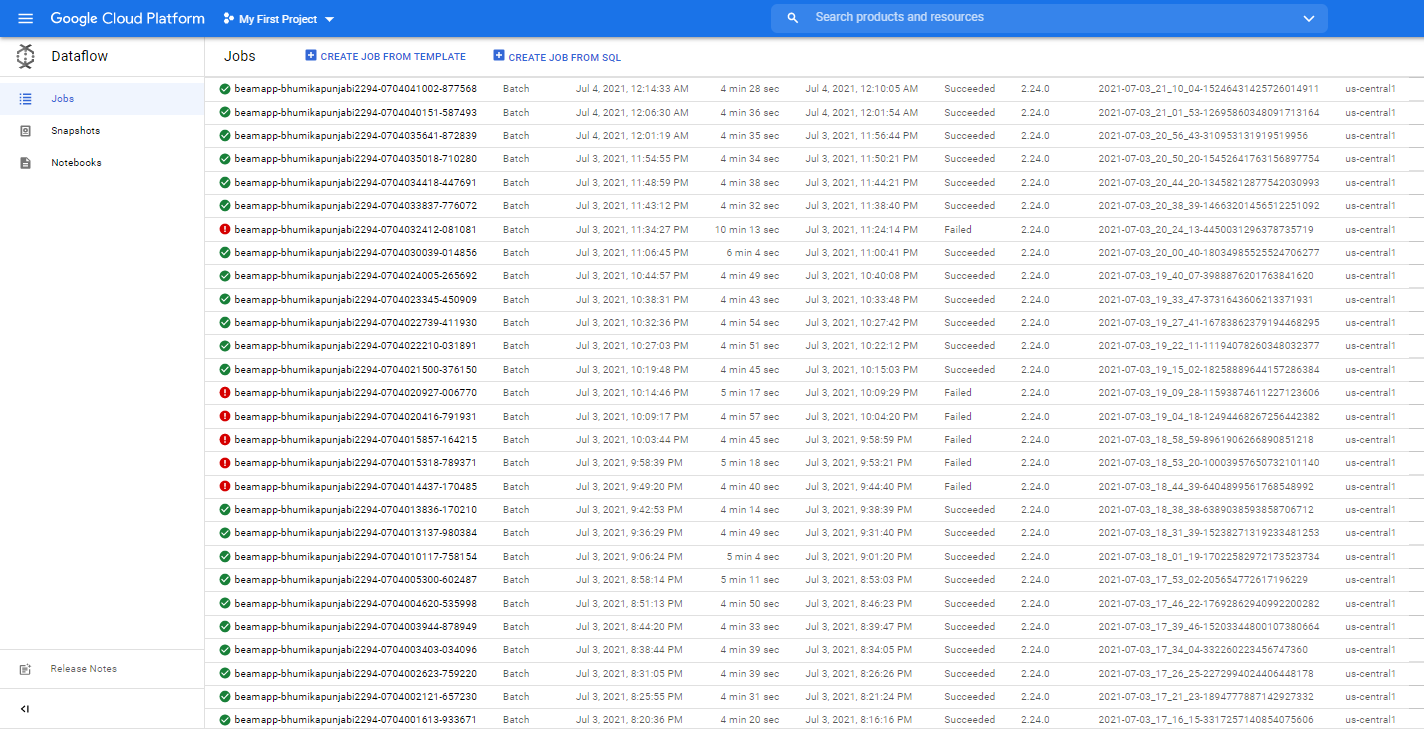
Pipeline to join Bigquery tables-



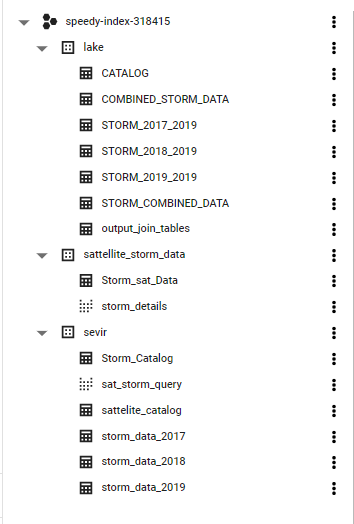
And this is how the Dataflow DAG looks like on the GCP console when you expand the ‘Inner join’ transform.



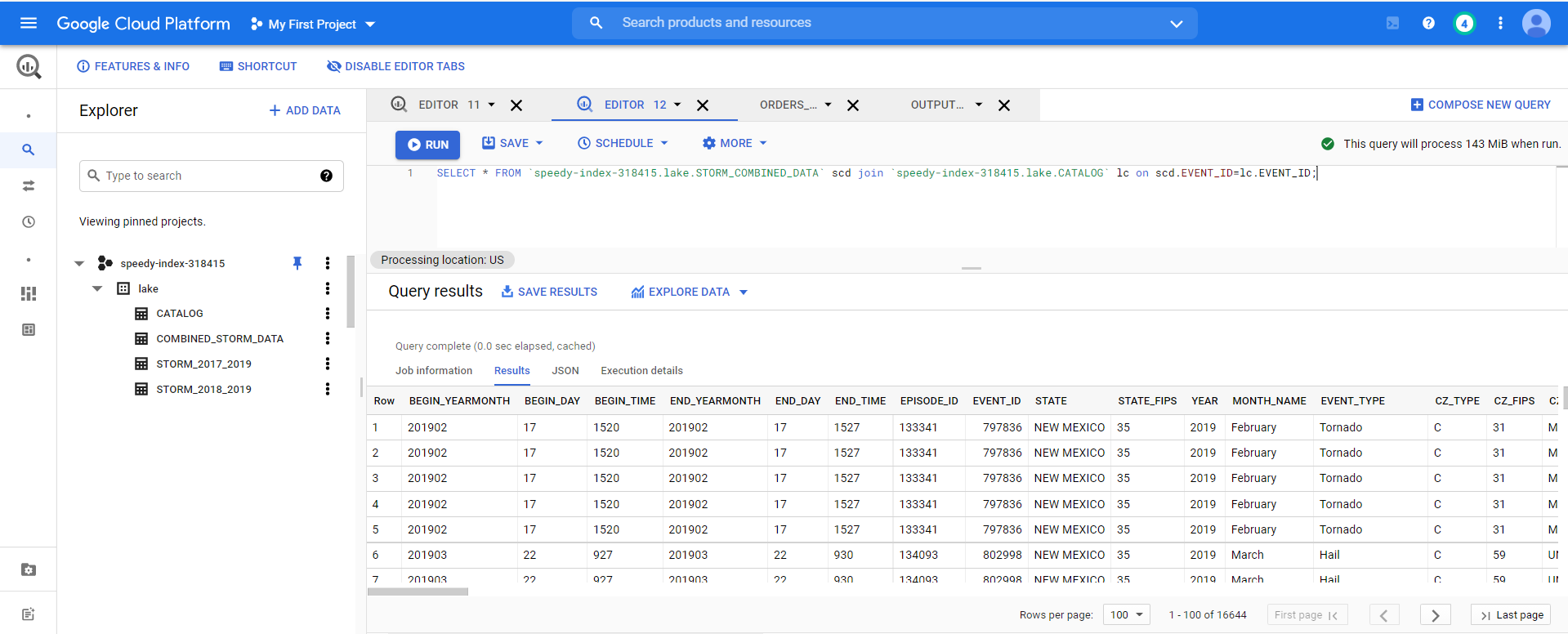
Dataflow logs(After running all pipelines)



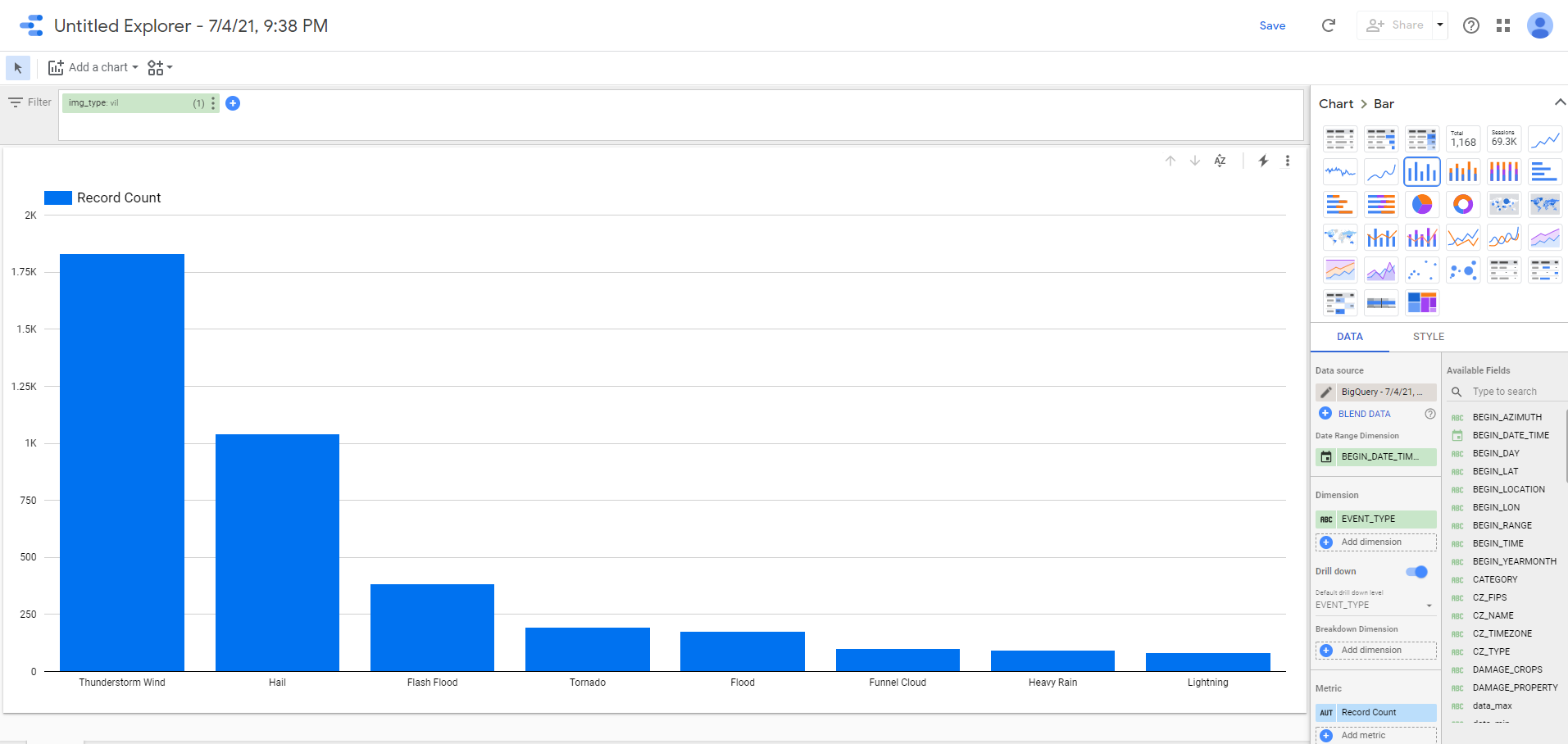
Once it is completed and succeeded, we can see results in the BigQuery

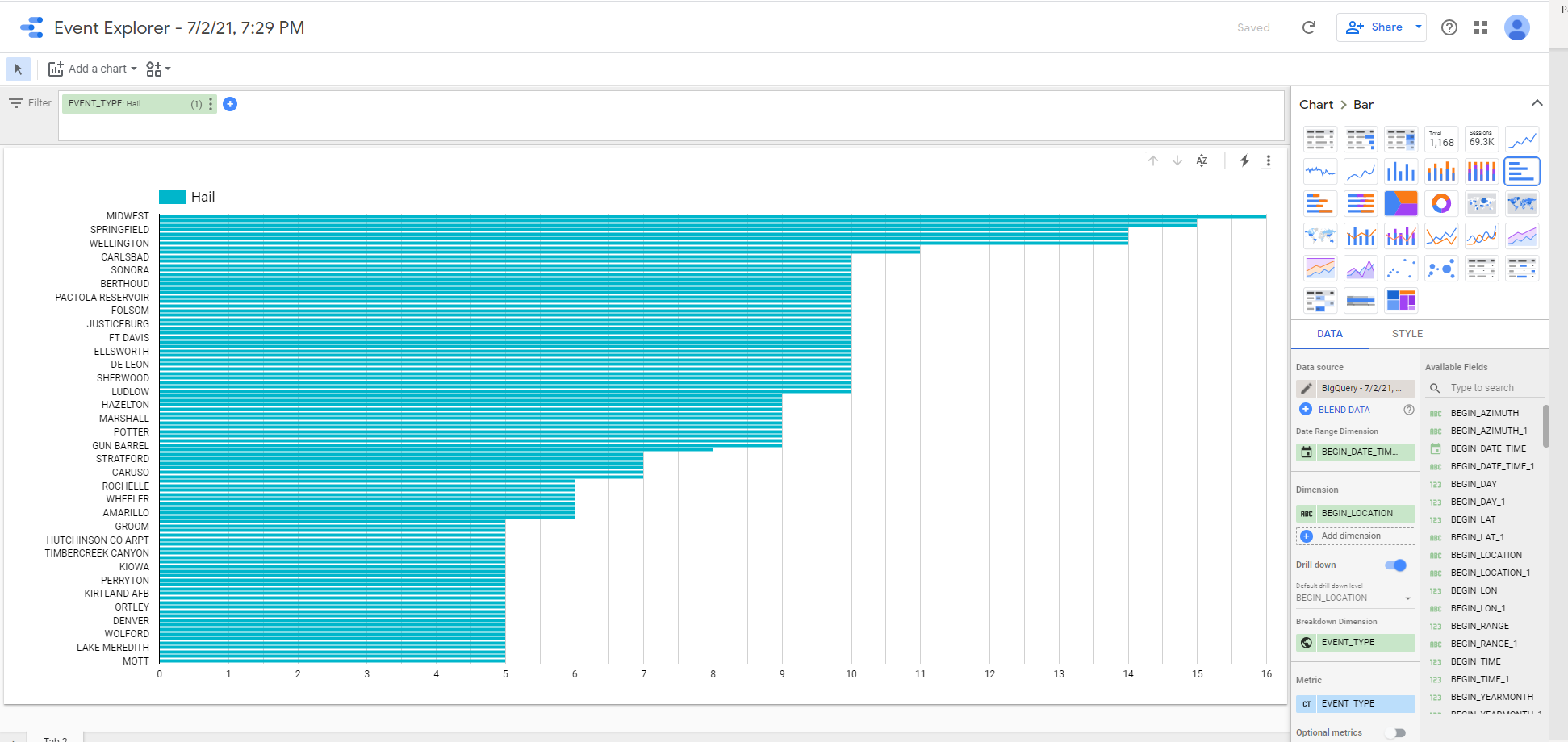


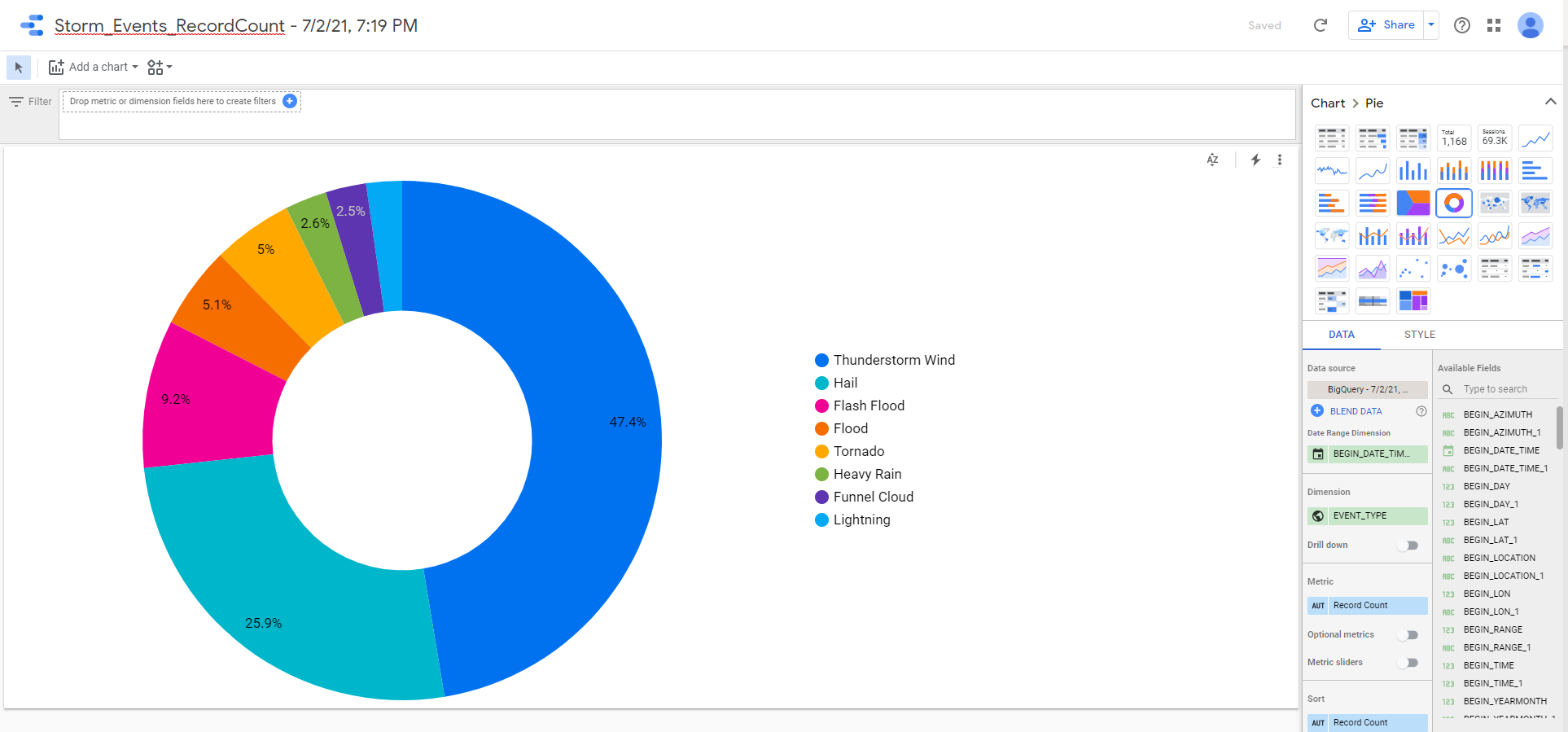
We can query out the data to get some insights.



Visualizations using Data Studio-





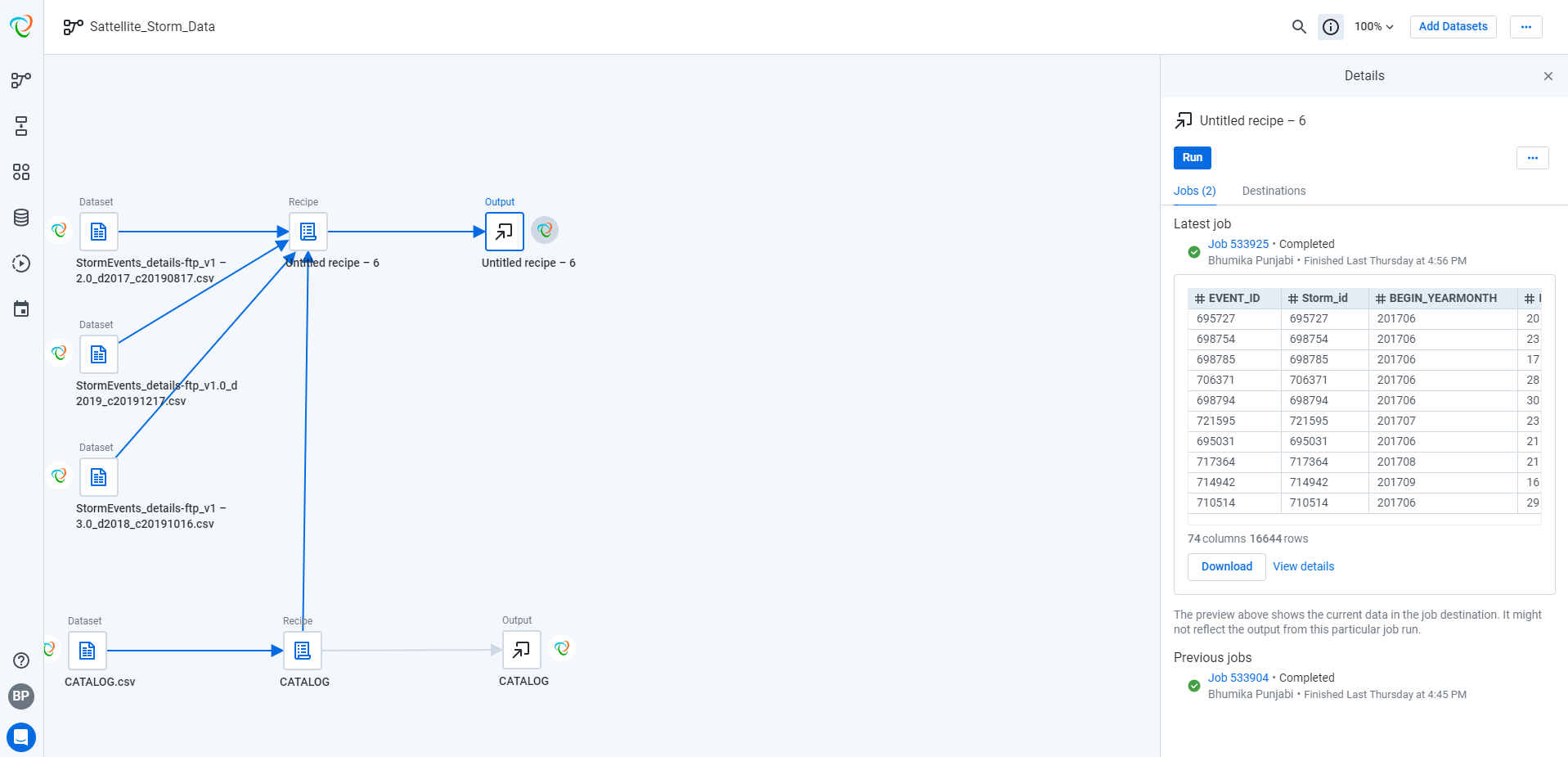


Link to google cloud-

https://console.cloud.google.com/bigquery?authuser=0&project=speedy-index-318415&ws=!1m0

Different Architecture

Preprocessing (Merging Data) using Trifacta Data pipeline-



Recipes

