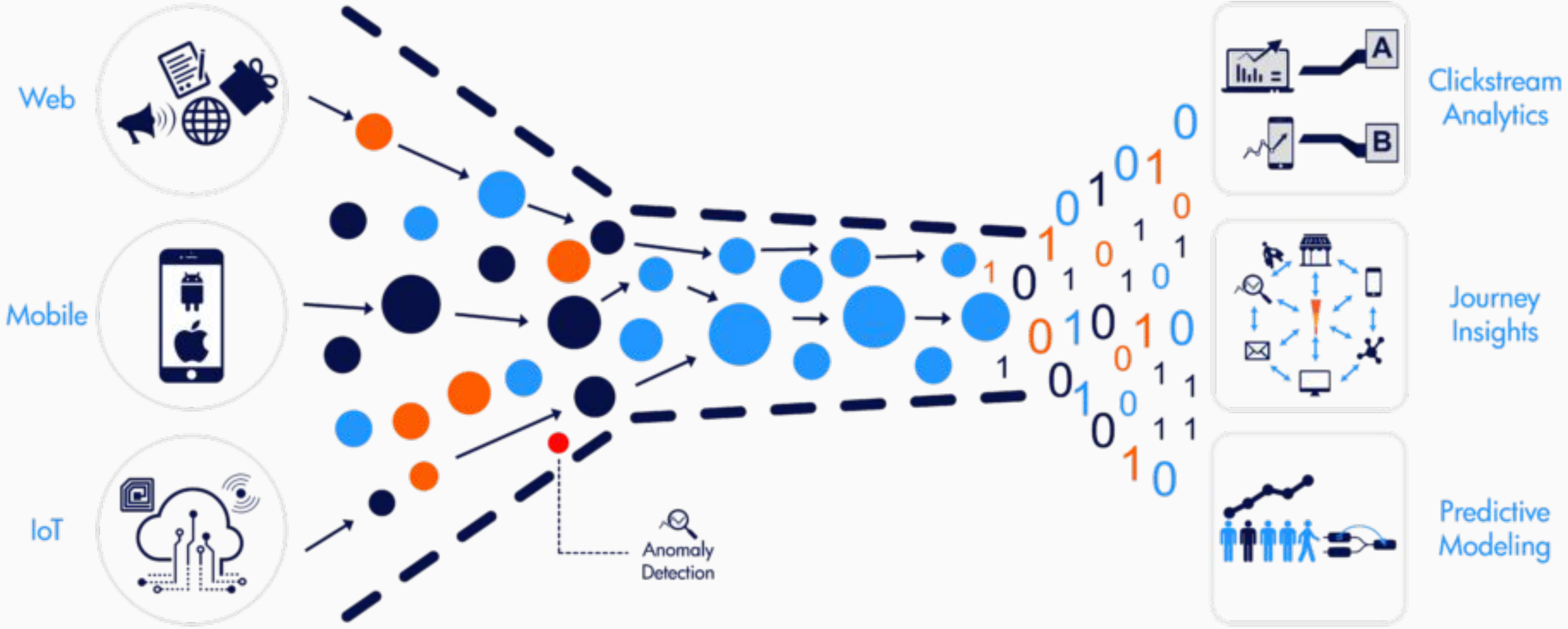
Globally Scalable Clickstream Analytics on Google Cloud

PyData Bangalore - Oct 2019

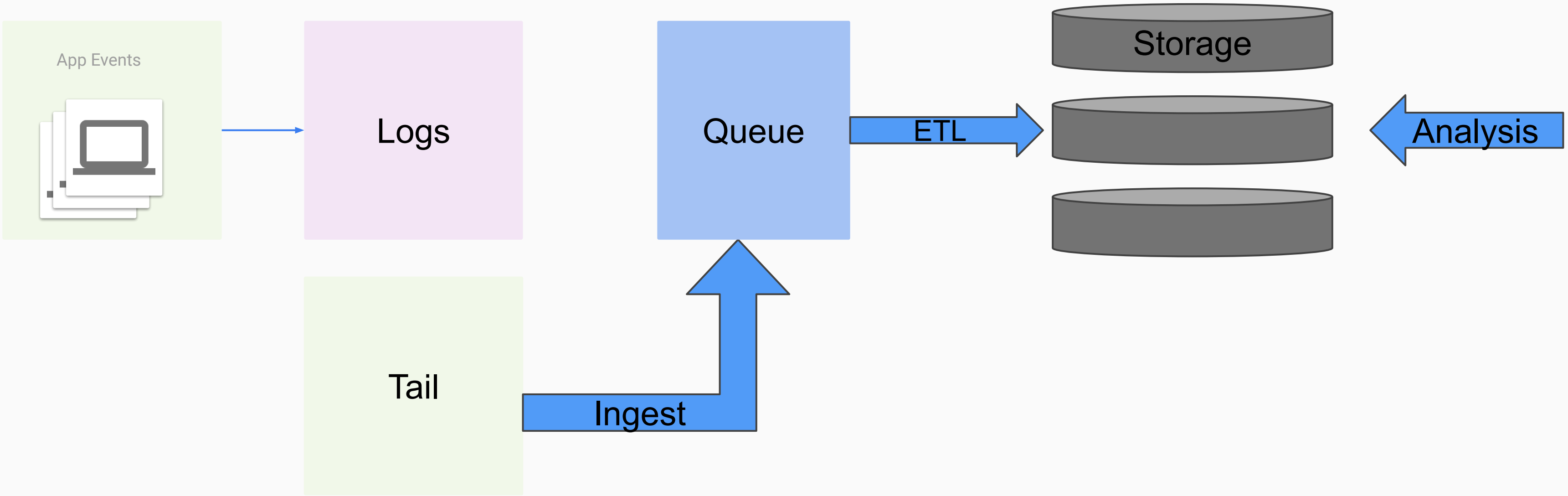
Ramjee Ganti

- Engineer
- Scaled Tech at Startups
- Entrepreneur
- Anti Sugar Advocate
- @gantir

What is Clickstream?



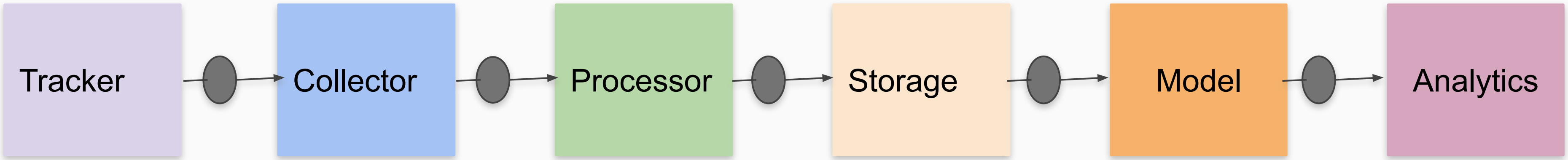
Logical Flow



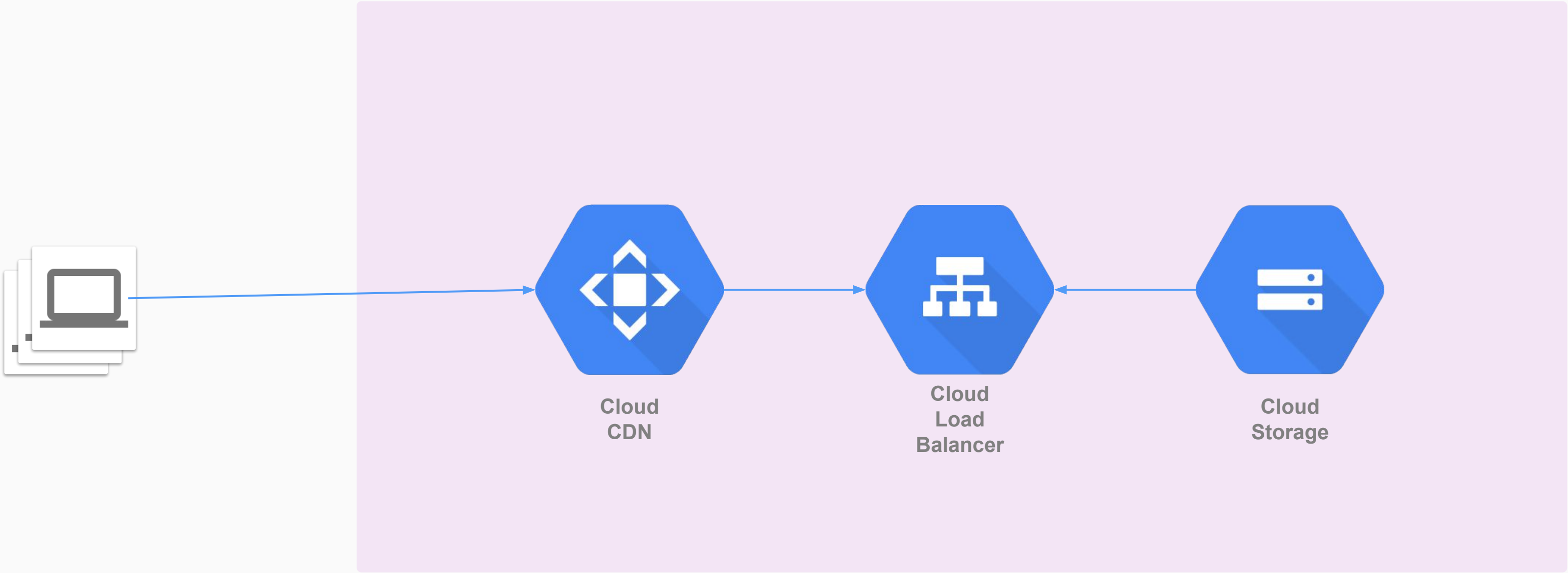
Concepts

- Tracking Protocol
- Versioned Schemas

Components



Tracker (Web Tracker)



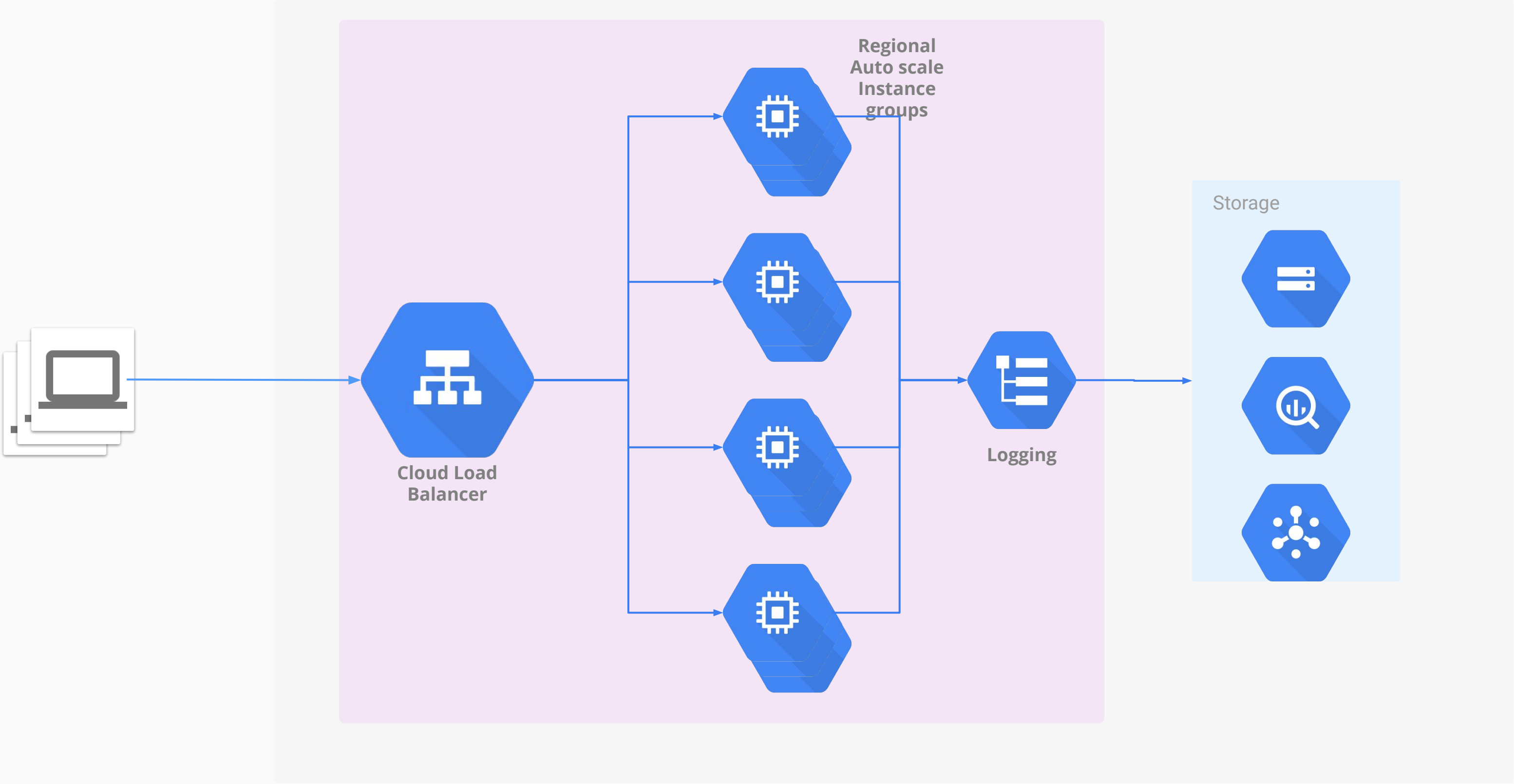
Collector

1. Only 1st party cookie
   1. Cloud CDN
2. 3rd party cookie
   1. GCLB + Instances
   2. GCLB + Kubernetes
   3. GAE
   4. Cloud Functions

Collector - Globally Scalable?

1. Only 1st party cookie
   a. Cloud CDN
2. 3rd party cookie
   a. GCLB + Instances
   b. GCLB + Kubernetes

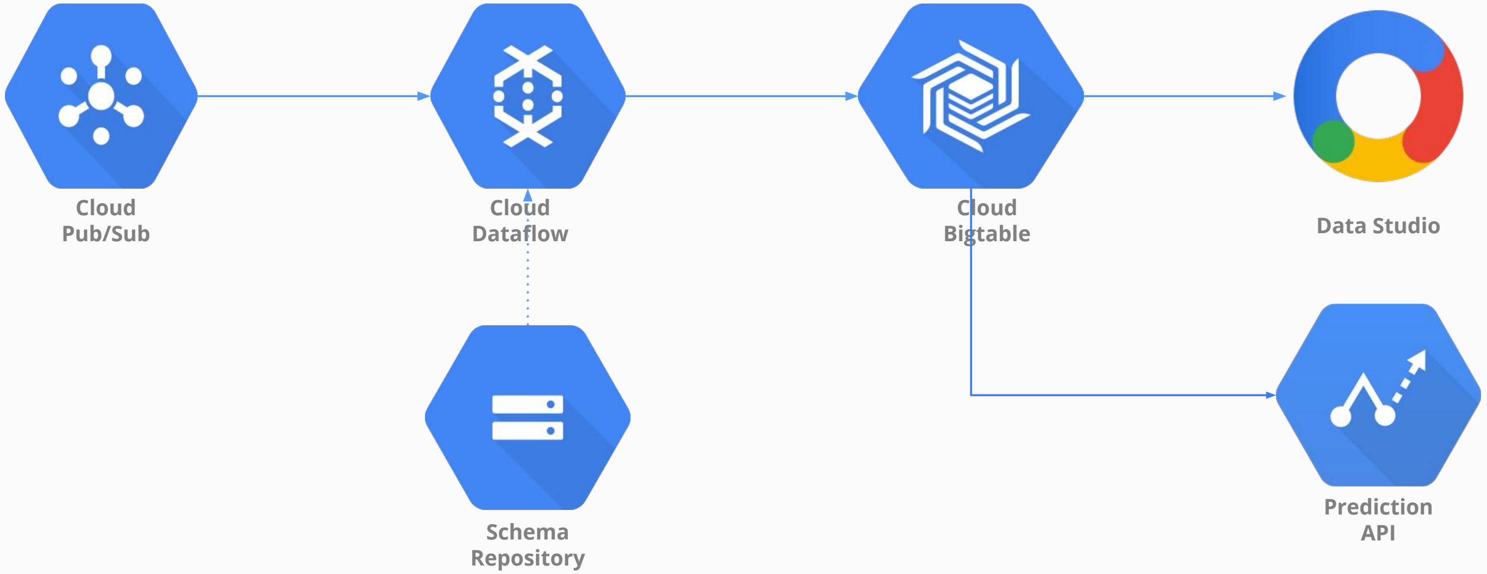
Collector


Regional
Auto scale
Instance
groups
Storage
Cloud Load
Balancer
Logging

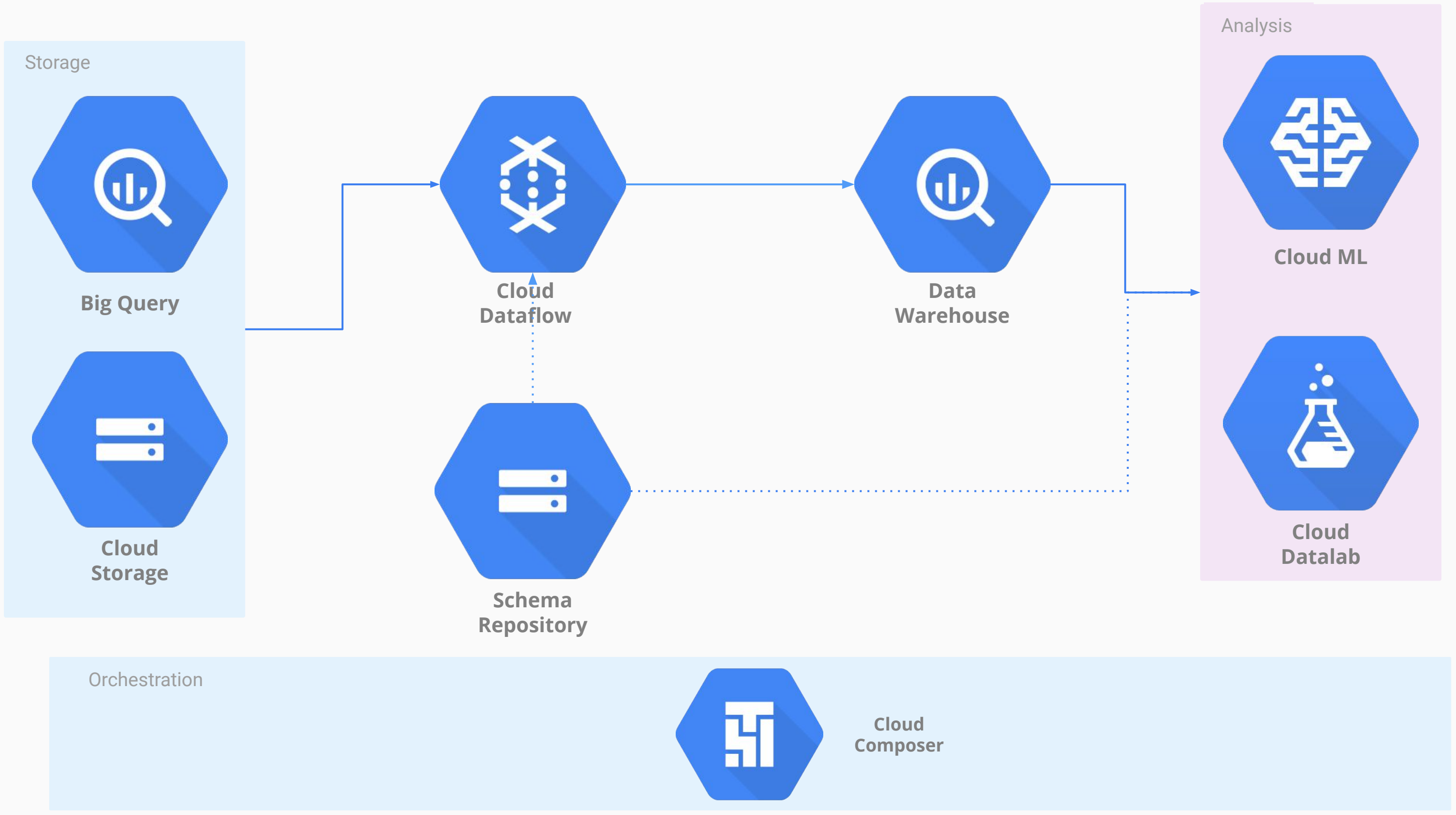
Apache Beam (Dataflow)

- Unified programming model for processing both batch and streaming data
- Cloud Dataflow is Google’s managed service on GCP.
- Infinitely scalable
- Only focus on the data and not the underlying infrastructure.
- Support for different runners (Spark, Flink, Hadoop MapReduce etc)

Real time


Cloud Pub/Sub
Cloud Dataflow
Cloud Bigtable
Data Studio
Schema Repository
Prediction API

Batch


Storage
Big Query
Cloud Storage
Cloud Dataflow
Schema Repository
Data Warehouse
Analysis
Cloud ML
Cloud Datalab
Orchestration
Cloud Composer

References

- Snowplow Analytics
- Google Cloud Solutions
- Streaming 101

Questions

Thanks

Annexure