Data Analytics and Insights Ecosystem on GCP

By Haider Lasne

Case Study

Main Theme

- Integrated Analytics and Insights ecosystem on GCP
- Transition to GCP

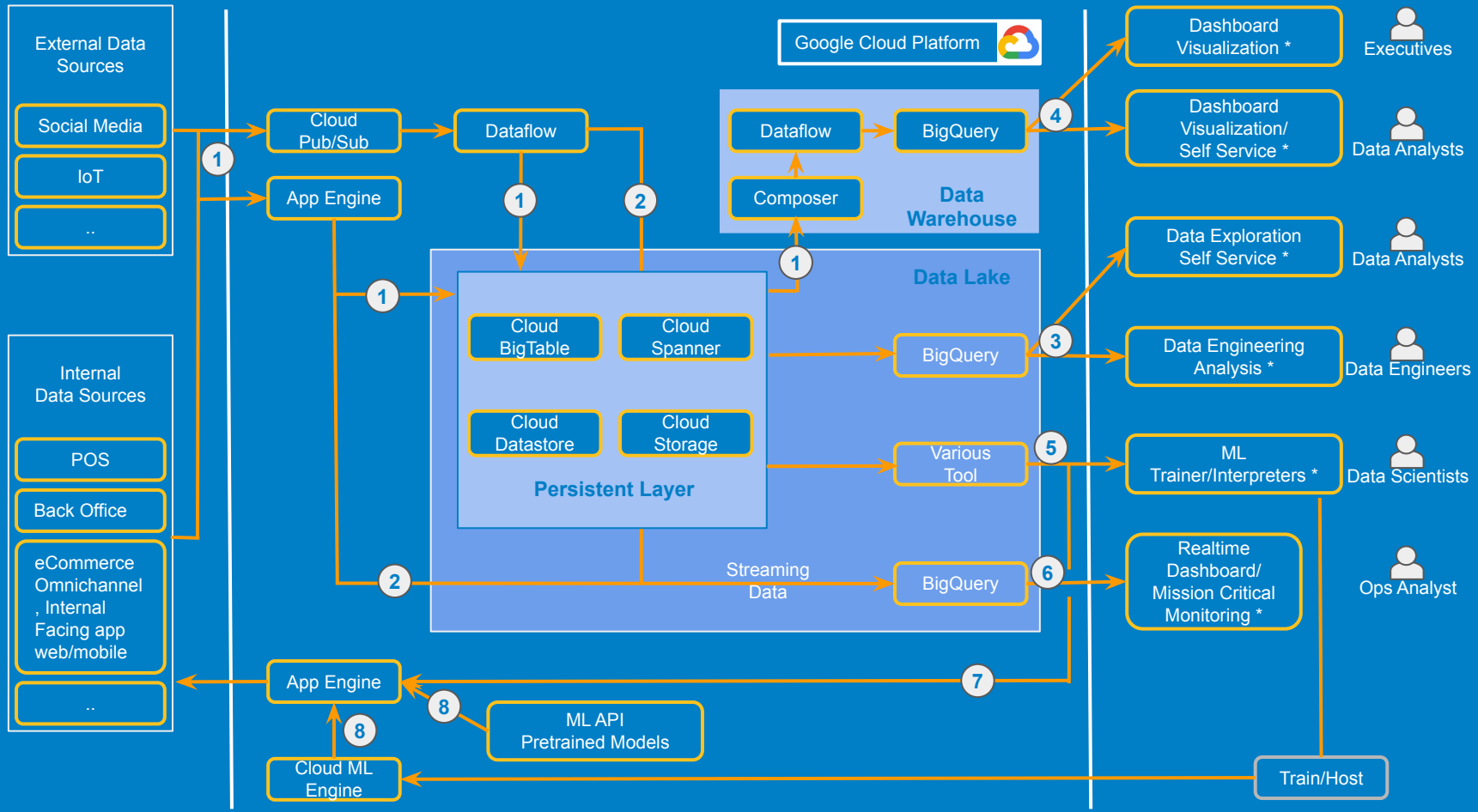
Primary Cloud Objective

- Enable Analytics Product with predictive analytics, Self Service and visualization capabilities.
- Enable Realtime and Offline Artificial Intelligence
  - Machine Learning based Data Quality
  - Machine Learning based Data Analysis,, Decision Making
- Elastic infrastructure (scale compute and storage on demand). Agile integration (connect with various systems with different technology stack).
- Build new systems on GCP (hybrid, multi cloud)
- Migrate existing systems to GCP. Few may stay back on current infrastructure.
  - Existing data engineering partially using Hadoop stack.

Additional Technical Considerations

- Data Ingestion
  - Streaming (IoT, POS, Financial Card, Social Media Data, eCommerce)
  - Batch (Finance, Vendor, etc.)
- Migrate to GCP database and data processing
  - SQL to Cloud SQL, Spanner
  - Migrate existing Hadoop stack to Dataproc, managed Hadoop on GCP
  - Move away from Hadoop stack with No Ops Dataflow, Bigtable
  - Migrate NoSQL (Casandra) to Bigtable for transaction systems, to Big Query for Analytics systems
- Cloud Data Lake and Datawarehouse
  - Storage: Cloud Storage, Bigtable, Big Query
  - Migrate from Hadoop HDFS
- Security at granular level @ dataset, fields, row
  - By Project separation (non integrated data and environments)
  - By functional role with IAM (by groups)
  - By data classification with database view
  - By data access logic with data access level functions
  - By Private Network, VPN and Service Policy

Data Analytics and Insights Ecosystem on GCP



Ecosystem on GCP Call Outs

1. The suggested generic ecosystem architecture should be refined and detailed further based on additional business need analysis.
2. Ecosystem Data Flow Themes: Streaming Data (1, 2, 6), Analytics and Insights (1, 4), Data Analysis and Engineering (Silver Data), Data Science (1,5,7,8)
3. Existing Hadoop stack can be migrated to Dataproc (managed hadoop, spark stack) and later transition to dataflow (ETL) and Bigtable.
4. GCP provides early access to new capabilities such as Composer which is Apache Airflow (workflow engine). For production rollout leverage respective open source software installing on Compute Engine.
5. Technology Selection Key Consideration: Storage throughput (query response) at baseline and load scenario, Privacy Data transport and storage.
6. Leverage preemptible machines such as for data proc for non prod or task not sensitive to time to lower the cost.
7. Application compute scalability can be addressed by GCP App Engine, Compute Engine, Kubernetes
8. Data Visualization, Self Service Analytics can be implemented with vendor supported software such as Tableau, MicroStrategy. (Viable open source option research in progress). Machine Learning ensemble and other analytics capability can be aggregated with thin UX layer.
9. Machine Learning based data quality can be implemented to identify data issue and data correct recommendations for example frequently data removal with TFIDF (term frequency–inverse document frequency), flagging data for categorizing duplicate,incorrect or suspicious entries. Scoring data for quality based on uniformity, distribution, outliers, etc.

Thanks!

Any Question or Feedback?

Find me at linkedin.com/in/haiderlasne

Scalable Distributed Systems Considerations

Design Principles

Availability	Performance	Reliability	Scalability	Manageability	Cost

Assumptions Not True

- Reliable, Homogeneous,Secured Network.
- Zero Latency
- Infinite Bandwidth
- Constant Topology
- No Transport Cost
- Transport cost is zero.

System Design

- Service oriented Architecture
- Redundancy
- Partitions
- Feature Flag

Fast and Scalable Data Access

- Caches: Global Cache, Distributed Cache
- Proxies
- Indexes
- Load Balancers
- Queues

Monitoring

- End-to-end functionality
- each service independently.
- As end user will access
- At a frequency to detect issues before customers
- Establish a baseline
- Monitor the monitoring systems.