LTIMindtree

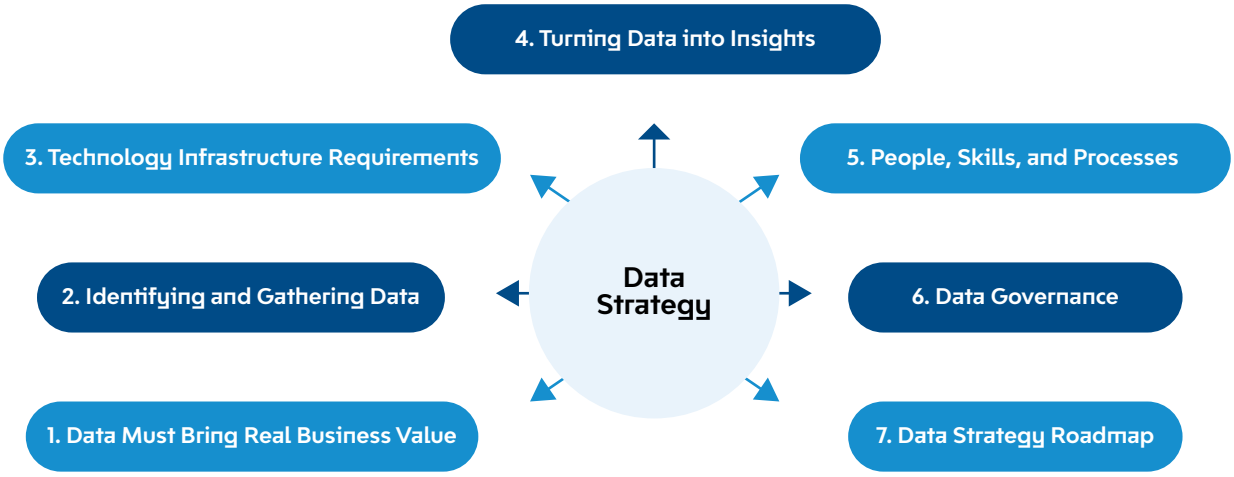
Point of View

Data Strategy for Google Cloud Platform

Author
Deepak Arora
Principal – Data Engineering BU, LTIMindtree

Companies are looking at leveraging data as a strategic asset to become data-driven, fostering data culture to improve decision-making, launching new products and services, and understanding customer journeys to provide personalized experiences and delight the customer. All of this is possible with a right and balanced data strategy to unleash data value. Data strategy is the first step of any digital transformational journey. It's a framework that allows you to generate business value through data and analytics.

There are seven key elements of any data strategy:



We will look at all these seven elements from the lens of Google Cloud Platform as to what the platform offers to build and become a data-driven organization.

Google Data Cloud provides a portfolio of solutions to implement the above elements of data strategy, from collecting the data to getting real-time actionable insights for businesses using Google Cloud-native tools, open source, third-party products, and solutions available on Google Marketplace. These services ensure that you can break down operational silos, build data pipelines, generate real-time insights, and make better business decisions.

Let Us Look at the Art of Possibilities in Each of the Areas

1. Data must bring real business value

The first and foremost step is to align data initiatives with business outcomes and strategy. With the help of in-house experts, SI, vendors, and Google professional services, identify the use cases you want to target first, instead of boiling the ocean, as they say. Based on the industry you are in, some of the use cases that can be targeted are, for example, anomaly and fraud detection, customer churn analytics, customer segmentation for personalized marketing, improved customer experience with real-time recommendations, data monetization, etc.

Instead of starting from scratch, Google provides several data analytics design patterns, which can be leveraged across industry segments to build analytics solutions, like anomaly detection for fraud detection, identifying emerging security threats, forecasting and recommendation systems, and many more.

2. Identifying and gathering data

Breaking data silos, data scattered in multiple places and formats, on-premise, or multiple clouds are some of the typical challenges you face when you embark on sourcing data into the Google Cloud. Google provides a wide variety of solutions and migration options to bring data into the Google Cloud. Tools like Pub Sub, Data Flow, and Data Fusion are some of the well-established solutions to integrate batch and real-time data into the Google Cloud.

Biglake is the Google Cloud answer to lake house centered around BigQuery with the interface to connect to any cloud object storage in a multi-cloud environment. It offers multiple open file formats and open-source processing engines, like Apache Spark or Beam.

Google's DataStream offering enables event-driven architectures, real-time data replication, and synchronization across heterogeneous databases and applications. Google also introduced the Spanner change stream, which can replicate data to BigQuery for real-time analytics and storage in Google Cloud Storage for compliance purposes.

Google has been focusing on simplifying the migration to the cloud with several migration services to accelerate and derisk the entire migration exercise. Database Migration service for Cloud SQL, live migrations from Apache HBase to Cloud Bigtable, BigQuery Migration Service for Teradata, to name a few.

Google recently launched BigQuery omni-cross cloud transfer service to integrate data from various cloud vendors. From a single interface, data citizens with different personas can load the data from AWS or Azure to BigQuery without any data pipelines.

3. Technology infrastructure requirements

Technology infrastructure is about choosing the right tool and architecture on how you get data into the organization, how you manage and govern the data, how you visualize and get insights from the data, how you build models, algorithms, machine learning engines, how you provision and share data with other systems, both internally and externally to the world.

The best part is that you do not have to worry about the majority of those infrastructure needs, as Google is going from strength to strength on their serverless capabilities.

Right from Data flow, Data fusion, Pub sub, cloud functions and BigQuery. You can look to build end-to-end data architecture solutions without worrying about managing any of the underlying infrastructure and straight away focus on building solutions that matter the most.

Leverage the Google Cloud Architecture Framework to adopt and implement best practices for optimizing cost using Finops to keep track on storage, compute databases, network, and cloud operations.

4. Turning data into insights

Customers are on their journey to generate value from data. They are looking at predictive analytics and taking actions in real-time. Google provides several tools like BigQuery, Vertex AI, and Databricks on Google Cloud, serverless spark service on Dataproc clusters to do the analytics job. You can bring data in BigQuery in real-time and use tools like Data Studio or Lookers to build dashboards or connect to Vertex AI to build, train, and deploy predictive ML models.

Vertex AI workbench-managed notebook capabilities can connect to any analytics spark engine, seamlessly integrate with big query and BiqQuery ML to explore data, and develop and train a ML model. These notebooks can be run as a pipeline using Vertex AI pipelines.

BigQuery BI engine, the Google in-memory analytics engine can integrate with many of the popular BI tools like Looker, Tableau, and Power BI without requiring any change to the BI tools.

BigQuery Omni provides a single pane of glass to analyze data sitting across Google, AWS, and Azure.

5. People, skills, and processes

Many organizations appoint a chief data officer, who is tasked to build the right team, set-up a modern and scalable architecture, establish processes for data sharing, and build the data culture in the organization. There is a whole new focus on democratizing the data and making it available for different personas in the organization, be it data engineer, machine learning engineer, data, or business analyst to collaborate, get value, and innovate from data.

Google has introduced Intelligent Data Fabric – Dataplex to provide the best of people, processes, and technology for a Data mesh architecture. Dataplex allows you to build domain-centric assets and teams and governance structure to orchestrate data provisioning across the enterprise. Data plex enables you to get integrated analytics experience to govern, secure, analyze and view the data from a single pane of glass with ability to automatically discover the data and make it discoverable through data catalog with a common model of policies and security of the data.

6. Data governance

Data governance is all about how you define trust in the data so that people believe in what they are looking at. Data security, data discoverability, and data quality are some of the key considerations when defining the data governance structure for an organization.

Google offers several tools to enable data governance in the organization, like Data catalog, DLP, and IAM. Data Catalog helps data discoverability, metadata management, and data level access controls using policy tags that allow separating sensitive data using Data loss prevention API. Big query provides fine-grained

access to sensitive columns using policy tag and define column-level security. AnalyticsHub data exchange allows you to securely exchange data sets, addressing challenges of data reliability using BigQuery and IAM security controls.

7. Data strategy roadmap

Roadmap is a fundamental tool that defines the execution methodology and explains how to implement and execute your data strategy. Laying out the plan for all the above pillars we talked about in a time horizon, so that we have a framework that we can manage, facilitate, and communicate on how we are going to make all these business changes happen. As they say, think big - start small, have a roadmap of the big picture while focusing the energy on delivering the business value at the earliest by showing quick wins, is the key to success for a remarkable journey.

Introducing Canvas Eureka

Canvas Eureka is one of the many accelerators built by LTIMindtree Data Practice to help customers migrate to the cloud. Canvas Eureka is a mindful automation framework and toolkit for GCP's Smart Analytics services, to forge end-to-end transformational change at scale and speed. Canvas Eureka takes an automation-first approach to address the core challenges of migrating legacy data solutions.

Solution highlights of Canvas Eureka -

- Analyzer tool of Canvas Eureka performs automated analysis of source systems and database objects to provide insights and create reports and drill-down dashboards.
- Meta Migrator automates migration of database schemas from DBMS databases to BigQuery/Cloud SQL.
- Data Migrator performs the actual migration from on-premise to BigQuery using spark API.
  SQL Converter – Automated conversion of source RDBMS/HiveQL queries/scripts to BigQuery queries.
- Data Validator provides an automated validation framework for data reconciliation from source to target environment.
- Canvas Eureka, with its automated dataflows, enables an optimized and seamless pathway to eliminate operational complexities, exploit cloud economics of scale, and activate AI/ML innovations.

Deepak Arora
Principal – Data Engineering BU, LTIMindtree

Deepak has more than 20 years of IT experience, including more than 13 years of experience in the Data Management space. He has extensive experience in collaborating with C-Suite, and Senior IT executives to develop data strategy roadmap and solutions. Deepak has helped several customers across the globe in embarking on their data modernization journey, setting up a data foundation platform and modern data architecture on Google Cloud platforms. Deepak is responsible for data architecture and technology consulting of strategic accounts at LTIMindtree. Deepak is also a certified Google Cloud Professional Data Engineer.