FOGHORN WHITEPAPER SERIES

A Resource Guide For Navigating The Cloud

Big Data = Big Opportunity on Google Cloud Platform

Best-in-breed tools to unlock the power of big data, to push future forward innovation further faster.

Google Cloud

foghornconsulting.com

Contents

Best-in-breed tools to unlock the power of big data, to push future forward innovation further faster.

INTRODUCTION

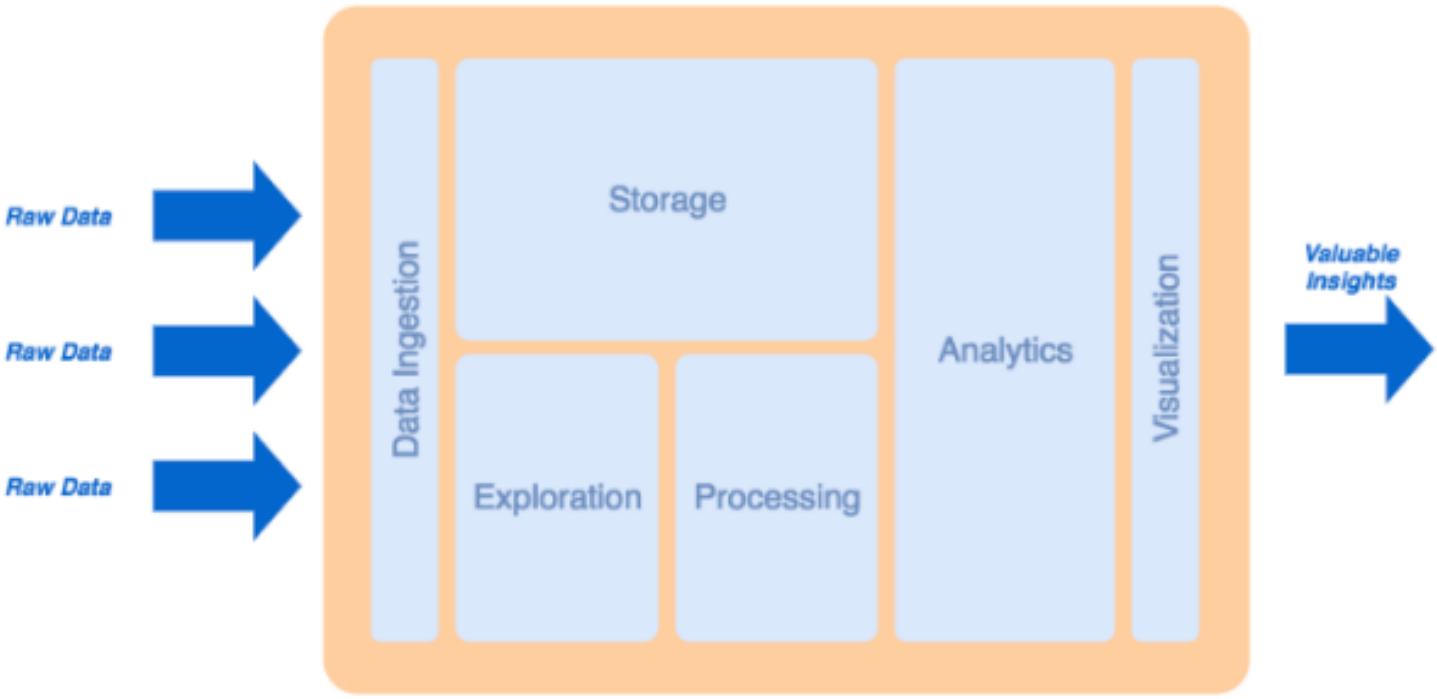
According to Forbes, 3.7 billion humans around the world use the internet, Google processes 40,000 searches a second (3.5 billion searches per day) **the most staggering stat is that 90% of the world's current data has been collected in just the past 2 years.**[1]

With the Internet of Things (IoT), sensors, social media and mobile device proliferation humans have never produced more bytes. While the modern technological universe has shown great expertise at collecting data, digesting it all hasn't always been simple for data scientists and business analysts. How do we take a bite of the bytes to digest this ocean of data to create meaningful insights, foster new discoveries and provide better service to customers to enhance the bottom line?

The good news is that cloud platforms have made paradigm shifting investments to harness the power of data. With unprecedented elastic compute power, the cloud of today is a perfect fit for data and the powerful insights it can provide. With world class tools that provide real time analytics enterprise can react in ways never thought possible before. As a leading Cloud Consultancy specializing in digital transformation, Foghorn fields many questions regarding Big Data. How can we combine disparate data sources and get a global view of our data across the organization to gain actionable insights to deliver better service for customers?

Google Cloud Platform, (born out of Google, which knows a thing or two about collecting and capitalizing on data) is often front and center in those conversations. Google Cloud Platform (GCP) continues to gain market share as their engineering, global infrastructure investments and their reputation of successful Business Intelligence (BI) outcomes increases demand for their platform. Garnter has named them a leader in Data Management Solutions for Analytics (DMSA). [2]

This whitepaper is an overview of GCP and Big Data. We will explore GCPs best-in-breed tools in regards to the five stages of data: capture, process, analyze and use. We will dive into GCPs industry leading tools, to inspire enterprise to grow their business with secure, compliant and performant Big Data on GCP.



High-level Big Data Platform Architecture

1. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/
2. https://cloud.google.com/gartner-magic-quadrant-for-dmsa/thank-you

CASE FOR BIG DATA IN THE CLOUD

For consumers, the term Big Data can drift into the world of jargon, and like the term Cloud, can become a victim of its own vagueness and obliqueness. Industry on the other hand is fully aware that like the Oakland A's Data Scientist Billy Bean as documented the book and film Moneyball, the large pools of Big Data are essential to beat the competition, find previously hidden gems and push future forward innovation further faster while driving down costs.

For decades companies have relied upon the on-premise enterprise data warehouse (EDW). The immense expense of maintaining these data centers meant the victor would be the company with the biggest IT spend. Beyond these steep costs the problem is that these operations were not elastic or scalable. With data expanding exponentially, the cloud provided the perfect ecosystem to right size the data warehouse with agility, real time scalability to keep pace with increased demands.

According to Gartner's Glossary, "Big data is high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."

From sensors on combines in soybean fields, to fitness trackers, to ecommerce shopping cart behaviors, Big Data holds the metrics for every business decision. Better market and customer intelligence leads to more sales and enhanced customer satisfaction. From HR to logistics, internal and operational efficiencies are enhanced. With big data built into the DNA of their product offering companies will be able to deliver huge value to stakeholders and customers alike.

3. https://www.domo.com/blog/the-man-behind-moneyball-the-billy-beane-story/
4. https://www.gartner.com/en/information-technology/glossary/big-data

WHY IS GCP A SOLID CHOICE FOR BIG DATA ANALYTICS IN THE CLOUD?

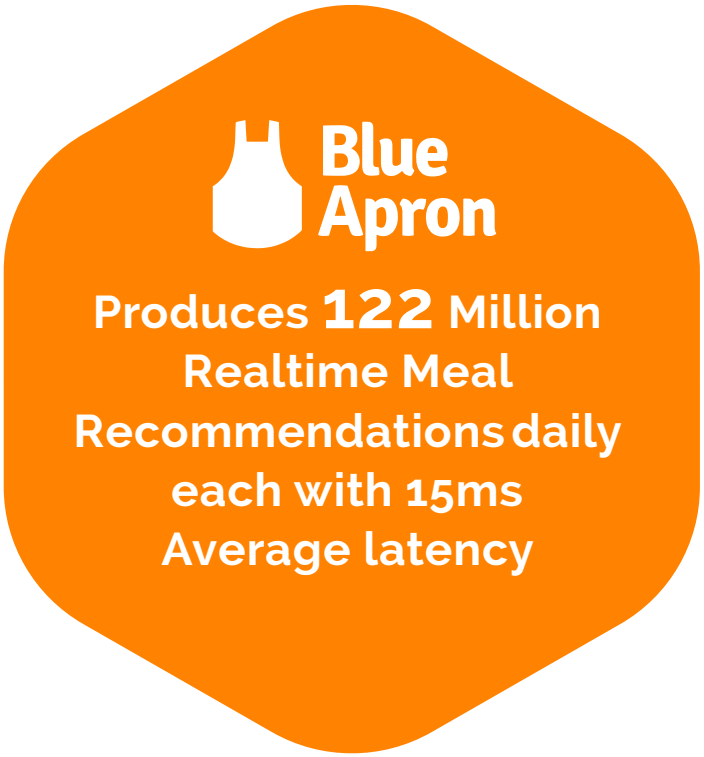
- Teams can shift focus from infrastructure to solving business problems.
- Utilize hybrid designs to honor existing infrastructure investments and take advantage GCP compute power and elasticity during usage spikes
- Stay competitive and on the leading edge of technology with access to new releases and open source features
- Have access to open source components that empower multi-cloud portability
- Achieve faster time to market with data applications
- Pay only for infrastructure you utilize
- Build performant architecture that drive down costs

BIG WINS WITH BIG DATA

Two companies who have transformed their business with Big Data on GCP

With the goal of feeding the world beautiful, nutritious, ethically sourced, proportioned meals, **Blue Apron** has become the leader in the emerging meal kit marketplace. Central to their success is their data intelligence regarding their customers and potential customers. Big Query and Data Flow have been essential with the creation of a virtuous cycle where an uptick in engagement results in better customer feedback where AI can tell the culinary team what meals people prefer.

As a start up Blue Apron found itself limited and frustrated due to vendor lock-in and have found a partnership with GCP to be a great comfort. "Write it once, run it everywhere, in our hybrid cloud and when Google is saying we are committed to an open cloud, that's very important to us," states Michael Collis, Technical Lead, ML and Personalization at Blue Apron.



Data has been a core part of Blue Apron's strategy. With fiber spread around the world, GCP's speed helps Blue Apron stay ahead of the competition. Blue Apron produces 122 million realtime meal recommendations everyday, each with 15 ms average latency. Data delivers better forecasting, and purchasing which in turn minimizes food waste and costs. Open source GCP tools have helped open up a more intimate dialogue with their fans to maximize customer lifecycle.

Operating in Malaysia, Indonesia, Thailand, Philippines, India and Japan, Malaysia based **Air Asia** is the largest low cost carrier in Asia. With 230 aircraft, each operating on average 14 hours a day and carrying over 19,000,000 passengers a year accuracy, efficiency and precision are central to Air Asia brand and growth plans. Having started in 2001 with just 2 aircraft, their story of growth has been astounding and pivots on their ability to deliver their services in the most cost efficient way possible. Central to this competitive advantage is their utilization of technology, specifically Google Cloud Platform and Big Data.

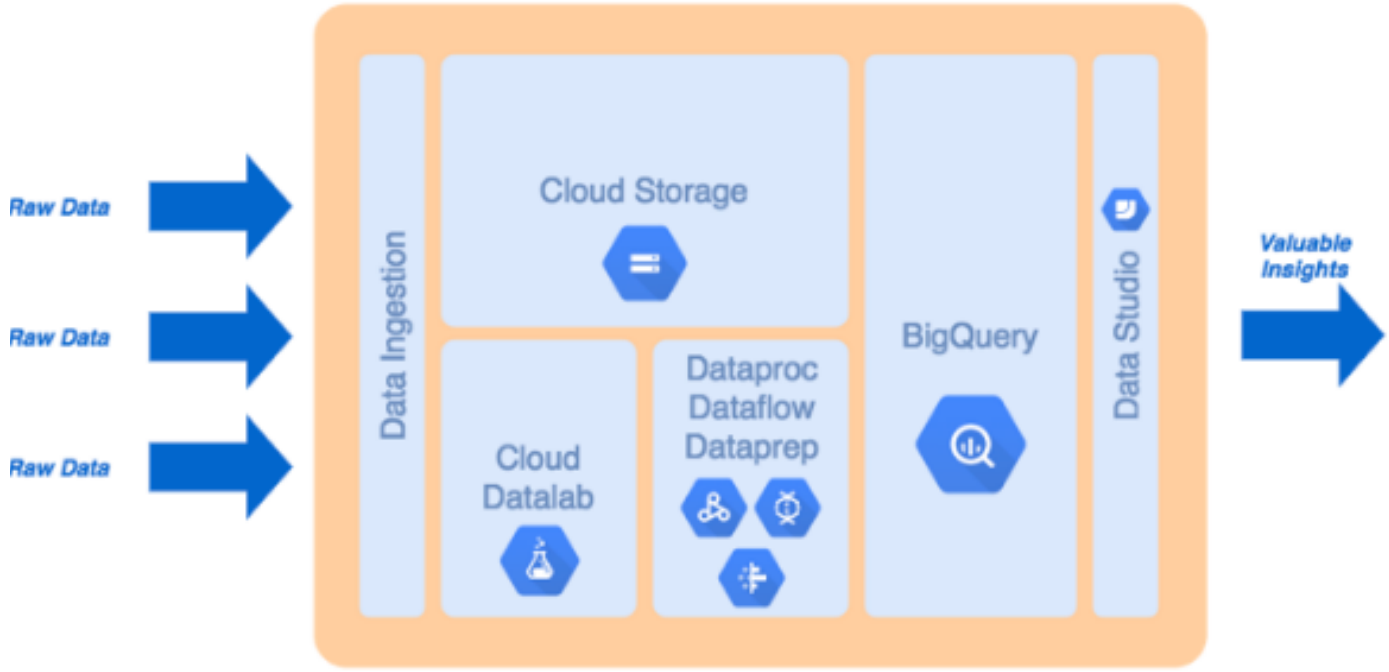
Over the past 18 years over 500,000,000 customers have flown on Air Asia's aircraft. 80% of their bookings have been made over the internet or on their mobile app. As efforts of digital transformation have become front and center within their organization 20% of their historical data has been ingested. With this intelligence Air Asia has created better demand forecasting and improved targeted marketing. Using BigQuery (as articulated by DataStudio) they know their customers better, and user patterns have emerged that Air Asia credits 2x conversion rates and 10% cost savings.

They also collect data on their operations, aircraft and their engines to maximize operational efficiency, reduce risk through predictive maintenance, and provide real-time weather forecasting and crew optimization. "BigQuery is a powerful tool that allows us to be scalable and be able to work faster and on the needs of our customers," shares Aireen Omar, Air Asia Deputy Group CEO and digital transformation lead.

5/6. Rethinking Big Data Analytics [Video] Retrieved October, 24, 2019, from https://www.youtube.com/watch?v=-VBGZsSUKlg .

BIG DATA TOOLBOX ON GCP

Cloud Consultants Foghorn and their savvy, experienced Solution Architects have deep experience helping organizations move from on-premise data centers into the cloud. With existing infrastructure investments, these migrations need not be an all-in affair. Google Cloud Platform has distinguished themselves by working well with Hybrid and Multi Cloud architectures. Before a company can migrate or lift and shift to GCP to harness the transformative capabilities of their Big Data tools a discovery of their tools is a great place to start.



High-level Big Data Platform Architecture on Google Cloud Platform

CLOUD STORAGE

Used by Spotify, Vimeo and CocaCola among many global titans of enterprise, Google Cloud Platform's unified object storage has three tiers (high, low and lowest) of high availability storage. With a single API across storage classes Google Cloud Storage is scalable to exabyte of data. Each storage class has very high availability, designed for 99.999999999% durability, with milisecond access to the first byte.

DATALAB

An interactive developer tool for analyzing and visualizing data, according to Google is designed to "get insights for raw data and explore, share and publish reports in a fast, simple and cost-effective way." Well known in data science universe, the service uses Jupyter notebooks format that enables analysts to create documents with live code and visualizations, thereby enhancing their power and potency within the organization. As Datalab is open sourced, developers can submit pull requests on GitHub. With this tool developers can make complex tasks simple and explore, transform, visualize and process data that is on BigQuery, Compute Engine and Cloud Storage.

DATAPROC

With many organizations running Apache Hadoop Distributed File System (HDFS) or Apache Spark, GCP realized that migrating to a new framework was not always viable once business critical Hadoop or Spark based data processing was part of the data fabric. Running these applications at scale can become costly as clusters used infrequently are billed the same as a persistent cluster. Cloud Dataproc enables a fast spin up of Spark or Hadoop cluster on demand, when needed, terminating clusters not in use.

DATAFLOW

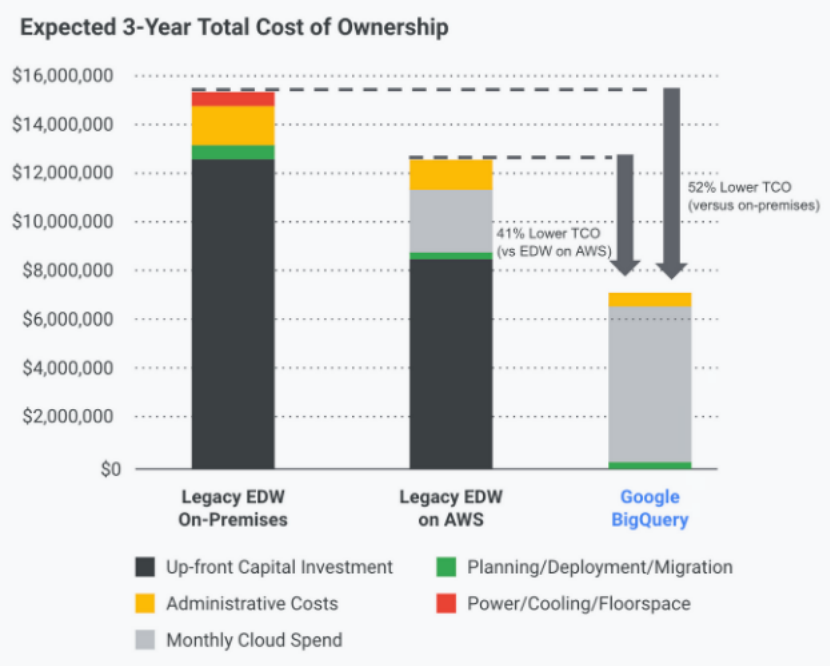
GCP's serverless, fully managed batch and stream processing tool, gives real-time filtering, aggregating, and streaming data for lightspeed high-volume analytics. Simplifying operational complexity and frustration, developers using Dataflow need not specify a cluster size or manage capacity. This zero-ops service frees up teams from having to consider cluster utilization and resources whereby they can focus on data transformation.

DATAPREP

Using Dataflow to scale automatically, Dataprep is an ideal tool to visually explore, clean and prepare data for analysis. Data agnostic Dataprep can help transform data of any size stored in CSV, JSON, or relational-table formats. With a browser based-UI clients do not need code deliver organized, clean, and powerful data into BigQuery.

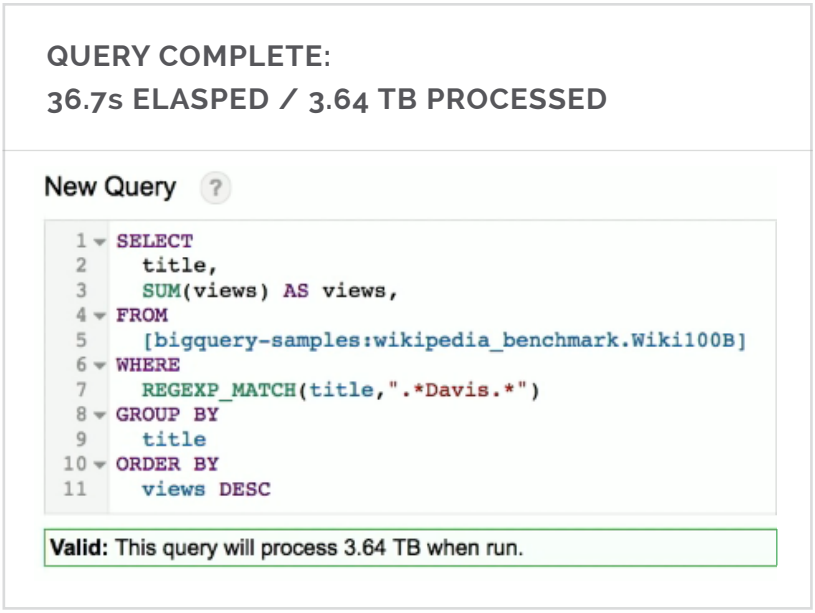
BIGQUERY

Google BigQuery is a cloud-based, fully managed, serverless enterprise data warehouse that supports analytics over petabyte-scale data. Not too long ago the analyzation data sets would take hours or even days. With the advent of GCP's Big Query a terabyte of data can be ingested in one second. A recent Enterprise Strategy Report states that using BigQuery as a data warehouse compared to an on premise data warehousing results in a 52% reduction in total cost of ownership (TCO).



When we discuss the secret sauce of Big Data on GCP, Big Query is front and center. With an easy to use UI end users only need to know SQL to begin using it. Forget indexes or VMs. Customers only pay for the data storage and the amount of data analyzed.

BigQuery provides seamless integration with most GCP services like Tensorflow, Cloud Dataflow, and Google Data Studio to name a few. With support of streaming ingestion as well as support for native machine learning (ML) capability, BigQuery is proving to be an indispensable tool for a growing number of enterprises.



Speed Data sets are Ingested, queried, and exported with blazing speeds.
Reliability The Google Cloud Platform provides always-on availability with goe replication provides constant uptime.
Security Utilizing Google Cloud Platforms Identity and access management (IAM) and encryption best practices, compliance across verticals is realized, including SOC, PCI, ISO 27001 and HIPAA.
Cost optimization Self scaling and returns resources automatically when ingestion is complete. If tables are not edited in 90 days, the storage price for that table automatically drops by 50%.

BIGQUERY ML

Whether is is uncovering ecommerce shopping cart patterns predictive performance models can deliver profitable insights. Google has always been on the frontiers of Machine Learning, and they have recently released this tool to integrate within their greater ecosystem.

THE THREE MODELS CURRENTLY SUPPORTED

> Linear regressions are used to predict the results of a continuous numeric variable, like income. Binary logistic regressions are used to predict the results of a categorical variable with two possible classes, such as whether a user will buy or not.
> Multinomial logistic regressions (or multiclass) are used to predict the result of a categorical variable with more than two classes.
> The smarter our datasets become that smarter they become. GCP and ML are in the early nings of this evolution and it's already proved to be exciting- at least that is what Foghorn's ML analysis predicted.

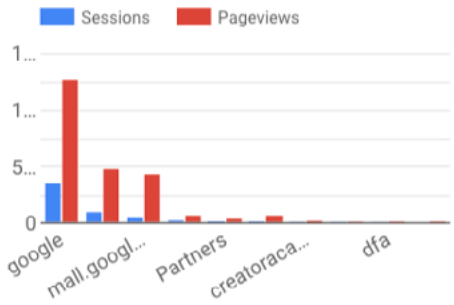
DATA STUDIO

The mouth of the data funnel to the cup of valuable insights, Data Studio unites all of a client's disparate data in one visually stunning and powerful place. Data scientist can tell stories about the data to translate to the rest of the organization to push forward, pivot or change course. The best part of Data Studio is the price of free. With over 150 integrations on top of the Google ecosystem Data Studio can provide holistic data digestion results for many enterprises.

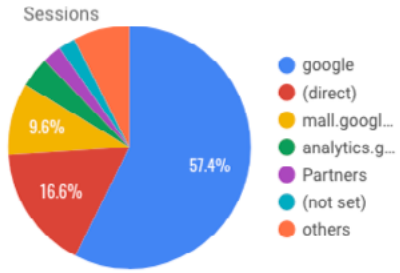
Storytelling with Data Simplified

- Connect data source
- Create visualizations by dragging and dropping data
- Share finished creation with just your internal team or the world.

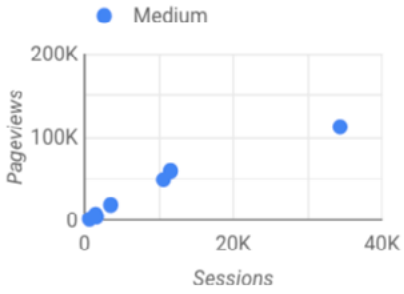
Geo map
Use geo dimensions to visualize your data in the real world



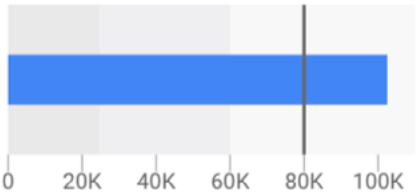
Bar chart
Compare metrics using vertical or horizontal bars



Pie chart
Quickly show relative differences in your data



Scatter chart
Discover correlations within a dimension



Bullet chart
See how well your data performs against target

Area chart
Discover correlations within a dimension

7. https://cloud.google.com/storage/docs/storage-classes
8. https://www.esg-global.com/the-economic-advantages-of-migrating-enterprise-data-warehouse-workloads-to-google-bigquery
9. https://datastudio.google.com/data

FUTURE IS HERE

Every aspect of modern business has or will be impacted by Big Data. From mom and pops to the Fortune 500 data collection, analysis, and interpretation will transform how we make decisions. With the intuitive design, global distribution, compliant, blazingly fast and well thought out tools Big Data produces big opportunities on Google Cloud Platform. Google's Big Data centerpiece BigQuery enables organizations to minimize cost and complexity traditionally associated with building and maintaining a fast, scalable, and resilient big data infrastructure. Customers who utilize GCP's best in class Big Data Platform leave costly and time consuming operations tasks like scalability, replication, protection, and recovery to Google. Instead of infrastructure management data teams can focus on gleaning new insights to improve the bottom line, make customers lives better and change the world.

From initial lift and shift to hybrid and multi cloud configurations, Foghorn Consulting has migrated, integrated and helped transform data sets for clients across verticals on Google Cloud Platform. Get in touch and one of our Solution Architects can deliver a roadmap for your journey to the promise of Big Data.

FOGHORN IS A GOOGLE CLOUD PLATFORM PARTNER

Google Cloud Partner

With the brains, heart, and muscle of Google behind them a growing number of enterprises are realizing innovation leaps with high-performance, high-security, and high-availability Google Cloud Platform. Whether part of a hybrid, multi-cloud or all-in strategy Google Cloud Platform (GCP) has world class tools and infrastructure to propel industry out of IT management and into IT enablement. Foghorn has the expertise to guide you on this journey, assisting your team to build, scale, automate and leave potential in rear view mirror.

COMPLIMENTARY GCP BIG DATA CONSULTATION

SCHEDULE YOURS TODAY

Foghorn Consulting was founded in 2008 with a mission to ensure that cloud computing initiatives deliver maximum value for its customers. Based in the Silicon Valley, Foghorn provides domain expertise in strategy, planning, execution and managed cloud services to high-growth and enterprise companies seeking a cloud partner. Our team of DevOps engineers, SRE's and certified cloud architects bring over 20 years of domain expertise to ensure your cloud initiatives are a success.

FOGHORN

330 Townsend St, Suite 202

San Francisco, CA 94107

foghornconsulting.com

info@foghornconsulting.com

650-963-0980