iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** XI **Month of publication:** November 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue XI, November 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Google Cloud Platform: A Powerful Big Data Analytics Cloud Platform

Mandeep Kumar
MCA Student, Department of Computer Science, Central University of Haryana, India

Abstract: ***Google Cloud Platform provides a powerful Big Data analytics cloud platform in the form of BigQuery, Cloud Dataflow, Google Cloud Dataproc, Cloud Datalab, Cloud Pub/Sub, and Google Genomics used by all types of organizations. There is no infrastructure to manage and you don't need a database administrator, so you can focus on analyzing data to find meaningful insights using familiar SQL and easily process big datasets like Apache Hadoop, Apache Spark , Apache Pig, and Apache Hive service at low cost. Google Cloud Platform giving you a powerful and complete data processing platform.***
Keyword: Google Cloud Platform, Big Data, BigQuery, Cloud Dataflow, Google Cloud Dataproc, Cloud Datalab, Cloud Pub/Sub, Google Genomics

I. INTRODUCTION

Google Cloud Platform provides Big Data services in the form of BigQuery, Cloud Dataflow, Google Cloud Dataproc, Cloud Datalab, Cloud Pub/Sub, and Google Genomics. It provides a powerful Big Data analytics Cloud platform.

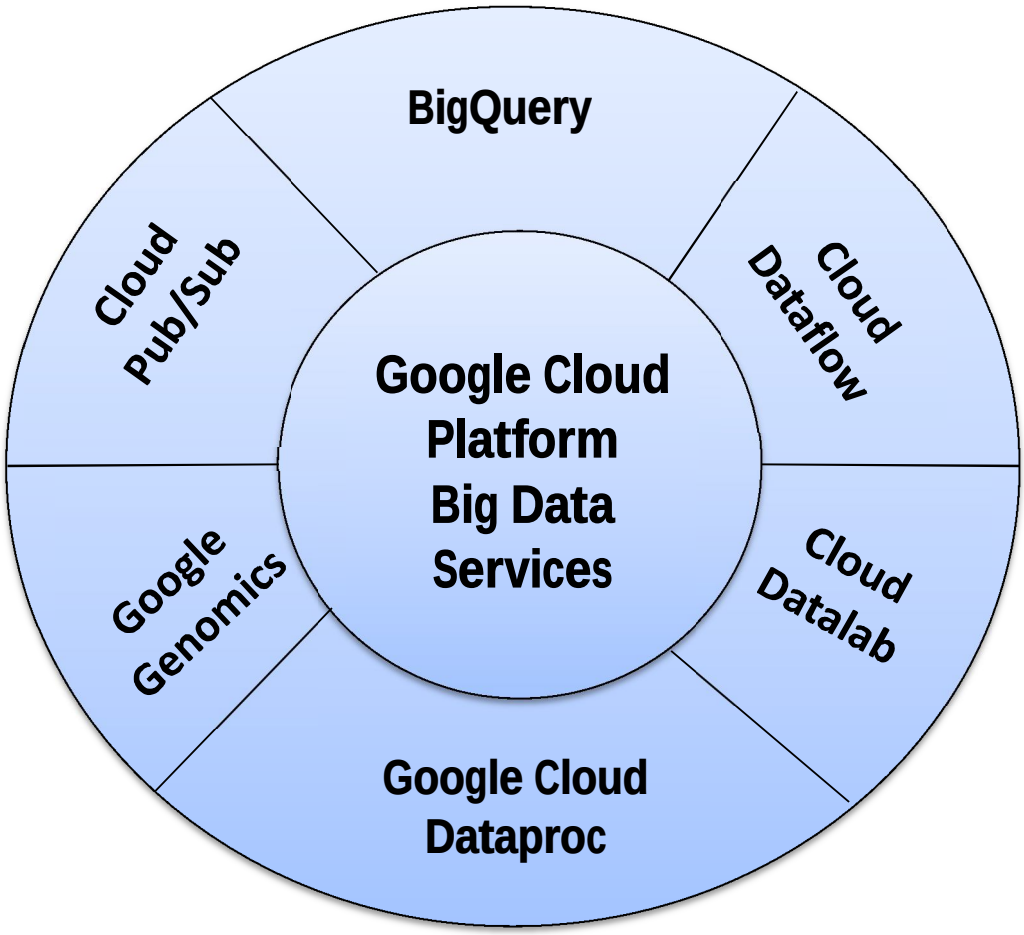


Fig 1: Google Cloud Platform Big Data Services

BigQuery is Google's fully managed, petabyte scale, low cost enterprise data warehouse for analytics. Cloud Dataflow is a unified

programming model and a managed service for developing and executing a wide range of data processing patterns including ETL, batch computation, and continuous computation. Cloud Dataflow frees you from operational tasks like resource management and performance optimization. Use Google Cloud Dataproc, an Apache Hadoop, Apache Spark , Apache Pig, and Apache Hive service, to easily process big datasets at low cost. Control your costs by quickly creating managed clusters of any size and turning them off when you're done. Cloud Dataproc integrates across Google Cloud Platform products, giving you a powerful and complete data processing platform.

Cloud Datalab is a powerful interactive tool created to explore, analyze and visualize data with a single click on Google Cloud Platform. It runs locally and optionally on Google Compute Engine and connects to multiple cloud services easily, so you can focus on exploring your data. Cloud Pub/Sub is a fully-managed real-time messaging service that allows you to send and receive messages between independent applications. You can leverage Cloud Pub/Sub's flexibility to decouple systems and components hosted on Google Cloud Platform or elsewhere on the Internet. Google Genomics helps the life science community organize the world's genomic information and make it accessible and useful. Big genomic data is her today, with petabytes rapidly growing toward exabytes. Through extensions to Google Cloud Platform, you can apply the same technologies that power Google Search and Maps to securely store, process, explore, and share large, complex datasets.

II. BIGQUERY

BigQuery is Google's fully managed, petabyte scale, low cost enterprise data warehouse for analytics. BigQuery is server less. There is no infrastructure to manage and you don't need a database administrator, so you can focus on analyzing data to find meaningful insights using familiar SQL. BigQuery is a powerful Big Data analytics platform used by all types of organizations, from startups to Fortune 500 companies.

A. *Benefits*

1) *Speed and Scale:* BigQuery can scan TB in seconds and PB in minutes. Load your data from Google Cloud Storage or Google Cloud Datastore, or stream it into BigQuery to enable real-time analysis of your data. With BigQuery you can easily scale your database from GBs to PBs.
2) *Incredible Pricing:* BigQuery separates the concepts of storage and compute, allowing you to scale and pay for each independently. It also gives you flexible pricing options to better suit your needs. You can either choose a pay-as-you-go model or a flat-rate monthly price for those who need cost predictability.
3) *Security and Reliability:* BigQuery automatically encrypts and replicates your data to ensure security, availability and durability. You can further protect your data with strong role-based ACLs that you configure and control using our Google Cloud Identity and Access Management system.
4) *Partnerships and Integrations:* Google Cloud Platform partners and 3rd party developers have developed multiple integrations with BigQuery so you can easily load, process, and make interactive visualizations of your data. Our partners include Looker, Tableau, Qlik, Talend, Google Analytics, SnapLogic and more.

B. *Features*

A fast, economic and fully managed data warehouse for large-scale data analytics.

1) *Flexible Data Ingestion:* Load your data from Google Cloud Storage or Google Cloud Datastore, or stream it into BigQuery at 100,000 rows per second to enable real time analysis of your data.
2) *Global Availability:* You have the option to store your BigQuery data in European locations while continuing to benefit from a fully managed service, now with the option of geographic data control, without low-level cluster maintenance headaches.
3) *Security and Permissions:* You have full control over who has access to the data stored in Google BigQuery. Shared datasets will not impact your cost or performance (those you share with pay for their own queries).
4) *Cost Controls:* BigQuery provides cost control mechanisms that enable you to cap your daily costs to an amount that you choose.
5) *Highly Available:* Transparent data replication in multiple geographies means your data is available and durable even in the case of extreme failure modes.
6) *Fully Integrated:* In addition to SQL queries, you can easily read and write data in BigQuery via Cloud Dataflow, Spark, and Hadoop.

7) *Connect with Google Products:* You can automatically export your data from Google Analytics Premium into BigQuery, visualize it using Google Data Studio and analyze datasets stored in Google Cloud Storage.

III. CLOUD DATAFLOW

Dataflow is a unified programming model and a managed service for developing and executing a wide range of data processing patterns including ETL, batch computation, and continuous computation. Cloud Dataflow frees you from operational tasks like resource management and performance optimization.

A. *Benefits*

1) *Fully Managed:* The managed service transparently handles resource lifetime and can dynamically provision resources to minimize latency while maintaining high utilization efficiency. Dataflow resources are allocated on-demand providing you with nearly limitless resource capacity to solve your big data processing challenges.
2) *Unified Programming Model:* Dataflow provides programming primitives such as powerful windowing and correctness controls that can be applied across both batch and stream based data sources. Dataflow effectively eliminates programming model switching cost between batch and continuous stream processing by enabling developers to express computational requirements regardless of data source.
3) *Integrated & Open Source:* Built upon services like Google Compute Engine, Dataflow is an operationally familiar compute environment that seamlessly integrates with Cloud Storage, Cloud Pub/Sub, Cloud Datastore, Cloud Bigtable, and BigQuery. The open source Java- and Python-based Cloud Dataflow SDKs enable developers to implement custom extensions and to extend Dataflow to alternate service environments.
4) *Partnerships & Integrations:* Google Cloud Platform partners and 3rd party developershave developed integrations with Dataflow to quickly and easily enable powerful data processing tasks of any size. Integrations are done with the open APIs provided by Dataflow.

B. *Dataflow Features*

It provides reliable execution for large-scale data processing.

1) *Resource Management*: Cloud Dataflow fully automates management of required processing resources. No more spinning up instances by hand.
2) *On Demand:* All resources are provided on demand, enabling you to scale to meet your business needs. No need to buy reserved compute instances.
3) *Intelligent Work Scheduling:* Automated and optimized work partitioning which can dynamically rebalance lagging work. No more chasing down “hot keys” or pre-processing your input data.
4) *Auto Scaling:* Horizontal auto scaling of worker resources to meet optimum throughput requirements results in better overall price-to-performance.
5) *Unified Programming Model:* The Dataflow API enables you to express MapReduce like operations, powerful data windowing, and fine grained correctness control regardless of data source.
6) *Open Source:* Developers wishing to extend the Dataflow programming model can fork and or submit pull requests on the Apache Beam SDKs. Dataflow pipelines can also run on alternate runtimes like Spark and Flink.
7) *Monitoring:* Integrated into the Google Cloud Platform Console, Cloud Dataflow provides statistics such as pipeline throughput and lag, as well as consolidated worker log inspection-all in near-real time.
8) *Integrated:* Integrates with Cloud Storage, Cloud Pub/Sub, Cloud Datastore, Cloud Bigtable, and BigQuery for seamless data processing. And can be extended to interact with others sources and sinks like Apache Kafka and HDFS.
9) *Reliable and Consistent Processing:* Cloud Dataflow provides built-in support for fault-tolerant execution that is consistent and correct regardless of data size, cluster size, processing pattern or pipeline complexity.

IV. GOOGLE CLOUD DATAPROC

Use Google Cloud Dataproc, an Apache Hadoop, Apache Spark , Apache Pig, and Apache Hive service, to easily process big datasets at low cost. Control your costs by quickly creating managed clusters of any size and turning them off when you're done.

Cloud Dataproc integrates across Google Cloud Platform products, giving you a powerful and complete data processing platform.

A. Benefits

1) *Fast & Scalable Data Processing:* Create Cloud Dataproc clusters quickly and resize them at any time —from three to hundreds of nodes—so you don't have to worry about your data pipelines outgrowing your clusters. With each cluster action taking less than 90 seconds on average, you have more time to focus on insights, with less time lost to infrastructure.
2) *Affordable Pricing:* Adopting Google Cloud Platform pricing principles, Cloud Dataproc has a low cost and an easy to understand price structure, based on actual use, measured by the minute. Also, Cloud Dataproc clusters can include lower-cost preemptible instances, giving you powerful clusters at an even lower total cost.
3) *Open Source Ecosystem:* The Spark and Hadoop ecosystem provides tools, libraries, and documentation that you can leverage with Cloud Dataproc. By offering frequently updated and native versions of Spark, Hadoop, Pig, and Hive, you can get started without needing to learn new tools or APIs, and you can move existing projects or ETL pipelines without redevelopment.
4) *Have You Considered?:* Cloud Platform can deliver even more scale, efficiency, and simplicity for key data processing and analysis scenarios. If you use Hive on Hadoop (or SparkSQL) you might consider Google BigQuery, an on-demand SQL analytics service with amazing performance. If you program data transformation pipelines with Spark or MapReduce, you may want to consider Google Cloud Dataflow, a fully-managed service that eliminates the busy work required by other tools and executes a wide range of data processing patterns, including ETL, batch, and streaming computation.

B. Cloud Dataroc Features

Google Cloud Dataproc is a managed Spark and Hadoop service that is fast, easy to use, and low cost.

1) *Automated Cluster Management:* Managed deployment, logging, and monitoring let you focus on your data, not on your cluster. Your clusters will be stable, scalable, and speedy.
2) *Resizable Clusters:* Clusters can be created and scaled quickly with a variety of virtual machine types, disk sizes, number of nodes, and networking options.
3) *Integrated:* Built-in integration with Cloud Storage, BigQuery, Bigtable, Stackdriver Logging, and Stackdriver Monitoring, giving you a complete and robust data platform.
4) *Versioning:* Image versioning allows you to switch between different versions of Apache Spark, Apache Hadoop, and other tools.
5) *Developer Tools:* Multiple ways to manage a cluster, including as easy-to-use Web UI, the Google Cloud SDK, RESTful APIs, and SSH access.
6) *Initialization Actions:* Run initialization actions to install or customize the settings and libraries you need when your cluster is created.
7) *Automatic or Manual Configuration:* Cloud Dataproc automatically configures hardware and software on clusters for you while also allowing for manual control.
8) *Flexible Virtual Machines:* Clusters can use custom machine types and preemptible virtual machines so they are the perfect size for your needs.

V. CLOUD DATALAB

Cloud Datalab is a powerful interactive tool created to explore, analyze and visualize data with a single click on Google Cloud Platform. It runs locally and optionally on Google Compute Engine and connects to multiple cloud services easily, so you can focus on exploring your data.

A. Benefits

1) *Integrate and Open Source:* Cloud Datalab is build on Jupyter (formerly IPython), which boasts a thriving ecosystem of modules and a robust knowledge base. Cloud Datalab enables analysis of your data on Google BigQuery, Google Compute Engine, and Google Cloud Storage using Python, SQL, and JavaScript (for BigQuery user-defined functions).
2) *Scalable:* Whether you're analyzing megabytes or gigabytes, Cloud Datalab has you covered. Once you are satisfied with your transformation and analysis models, deploy them to BigQuery with the click of a button.
3) *Data Management and Visualization:* Use Cloud Datalab to gain insight from your data. Interactively explore, transform, analyze, and visualize your data using BigQuery, Cloud Storage and Python. Build machine learning models using TensorFlow.

4) *Cloud Datalab Features:* An easy to use interactive tool for large-scale data exploration, analysis, and visualization.
5) *Integrated:* Cloud Datalab handles authentication, cloud computation out-of-the-box and is integrated with BigQuery, Compute Engine, and Cloud Storage.
6) *Multi-Language Support:* Cloud Datalab currently supports Python, SQL and JavaScript (for BigQuery user-defined functions).
7) *Notebook Format:* Cloud Datalab combines code, documentation, results and visualizations together in an intuitive notebook format.
8) *Pay-per-use Pricing:* Only pay for the cloud resources you use; Google Compute Engine VMs, BigQuery, and any additional resources you decide to use, such as Cloud Storage.
9) *Interactive Data Visualization:* Use Google Charts or matplotlib for easy visualizations.
10) *Machine Learning:* Supports TensorFlow-based deep ML models in addition to scikit-learn.
11) *Open Source:* Developers wishing to extend Cloud Datalab can fork and/or submit pull request on the GitHub hosted project.
12) *Custom Deployment:* Specify your minimum VM requirements, the network host, and more.
13) *IPython Support:* Cloud Datalab is based on Jupyter (formerly IPython) so you can use a large number of existing packages for statistics, machine learning, etc.

VI. CLOUD PUB/SUB

Cloud Pub/Sub is a fully-managed real-time messaging service that allows you to send and receive messages between independent applications. You can leverage Cloud Pub/Sub's flexibility to decouple systems and components hosted on Google Cloud Platform or elsewhere on the Internet. By building on the same technology Google uses, Cloud Pub/Sub is designed to provide "at least once" delivery at low latency with on-demand scalability to 1 million messages per second (and beyond).

A. *Benefits*

1) *Connect Anything to Everything:* Use Cloud Pub/Sub to publish and subscribe to data from multiple sources, then use Google Cloud Dataflow to understand your data, all in real time. Use Cloud Pub/Sub to reduce dependencies between components of distributed applications. Cloud Pub/Sub is the same messaging technology used by many of Google's apps, from Ads to Gmail.
2) *Push and Pull:* Cloud Pub/Sub is designed for quick integration with systems hosted on the Google Cloud Platform or elsewhere, whether you need one-to-one, one-to-many, or many-to-many communication, with push or pull delivery.
3) *Guaranteed Delivery:* Cloud Pub/Sub is designed to provide "at least once" delivery by storing copies of messages in multiple zones to ensure that subscribers can receive messages as swiftly as possible. All message data is encrypted and protected on the write and at rest.
4) *Global and Scalable:* Cloud Pub/Sub is fully managed and global by design, automatically taking advantage of dedicated resources in every Google Cloud Platform region to ensure high-availability without degrading latency even under heavy load.

B. *Cloud Pub/Sub Features*

1) *High Scalable:* Any customer can send up to 10,000 messages per second, by default – and millions per second and beyond, upon request.
2) *Push and Pull Delivery:* Subscribers have flexible delivery options, whether they are accessible from the Internet or behind a firewall.
3) *Encryption:* Encryption of all message data on the wire and at rest provides data security and protection.
4) *Replicated Storage:* Designed to provide "at least once" message delivery by storing every message on multiple servers in multiple zones.
5) *Message Queue:* Build a highly scalable queue of messages using a single topic and subscription to support a one-to-one communication pattern.
6) *End-to-End Acknowledgement:* Building reliable applications is easier with explicit application-level acknowledgements.
7) *Fan-out:* Publish messages to a topic once, and multiple subscribers receive copies to support one-to-many or many-to-many communication patterns.
8) *REST API:* Simple, stateless interface using JSON messages with API libraries in many programming languages.

VII. GOOGLE GENOMICS

Google Genomics helps the life science community organize the world's genomic information and make it accessible and useful.

Big genomic data is her today, with petabytes rapidly growing toward exabytes. Through extensions to Google Cloud Platform, you can apply the same technologies that power Google Search and Maps to securely store, process, explore, and share large, complex datasets.

A. Benefits

1) *Get Results Sooner:* Query the complete genomic information of large research projects in seconds. Process as many genomes and experiments as you like in parallel.
2) *Scale to Power Any Project:* Whether you are working with one genome or one million, Google Genomics provides access to the power and flexibility you need to advance your work.
3) *Open and Interoperable:* Google Genomics supports open industry standards, including those developed by the Global Alliance for Genomics and Health, so you can share your tools and data with your group, collaborators, or the broader community, if and when you choose.
4) *Reliable Information Security:* Google's infrastructure provides reliable information security that can meet or exceed the requirements of HIPAA and protected health information.

B. Google Genomics Features

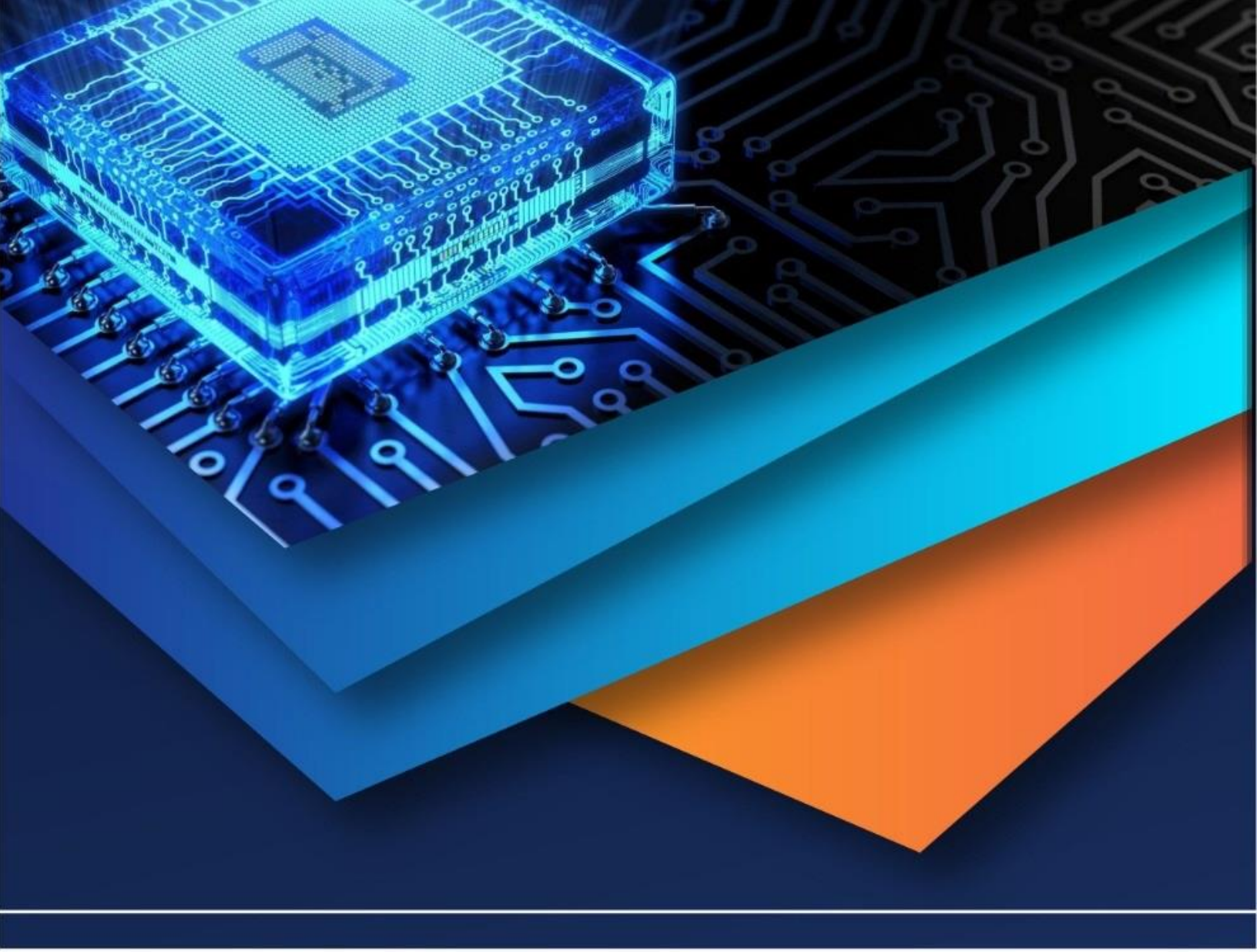
1) *Interoperability:* Our implementation of the open standard from the Global Alliance for Genomics and Health is interoperable across multiple genome repositories and it's backed by Google technologies like Bigtable and Spanner.
2) *Fully Integrated:* Google's Cloud infrastructure for your bioinformatics needs, including fast virtual machines, scalable storage, and a choice of fully managed SQL and NoSQL databases like Bigtable and Datastore.
3) *Real-time Data Processing:* Genomic data processing and analysis in real time with BigQuery, in literate programming style with Cloud Datalab, in batch with GATK on Google Genomics, with Apache Spark or Cloud Dataflow, or with a Grid Engine cluster.
4) *High Scalability:* You can load up petabytes of sequence reads, variants, references, and annotations, and process them all efficiently.

VIII. CONCLUSION

Google Cloud Platform provides a powerful Big Data analytics platform used by all types of organizations. Google Cloud Platform provides Big Data capabilities in the form of BigQuery, Cloud Dataflow, Google Cloud Dataproc, Cloud Datalab, Cloud Pub/Sub, and Google Genomics. It easily process big datasets like Apache Hadoop, Apache Spark, Apache Pig, and Apache Hive services at low cost. It runs locally and optionally on Google Compute Engine and connects to multiple cloud services easily, so you can focus on exploring your data. Google Cloud Platform giving you a powerful and complete data processing platform.

REFERENCES

[1] https://cloud.google.com/bigquery/
[2] https://cloud.google.com/dataflow/
[3] https://cloud.google.com/dataproc/
[4] https://cloud.google.com/datalab/
[5] https://cloud.google.com/pubsub/
[6] https://cloud.google.com/genomics/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY