ATSCALE

How to Maximize your Business Value with AtScale and Google BigQuery

Contents

Introduction

Big data is mainly defined by the size of a data set that it encapsulates. These data sets are ordinarily large in scale, measuring tens of terabytes and generally going over the petabyte scale. Originally known as very large databases and frequently managed using database management systems (DBMS) such as DB2, Teradata and Oracle, one of the main characteristics of big data is that it can be broken down into three different types of sets, structured, semi and unstructured. In this whitepaper, we will provide the information you need to understand how to cost effectively analyze big data on the cloud, through AtScale, using your favorite BI tools without having to worry about data movement or enforced proprietary formats, ultimately allowing you to reduce cost and increase performance.

Due to the indispensable nature of big data and analytics, in an effort to make it more accessible to the users, many organizations have procured and are running data warehouses as a central repository for all their data from single or multiple sources. This is a costly and complicated process. In the first implementations, a data warehouse was a new concept, data had to be extracted from mainframes and delivered to developers who would transform the data into a format users were able to consume using a SQL interface. While this was powerful, it was not user friendly.

The first generation of a semantic layer allowed IT to map complex data into familiar business terms, while enabling users to access data through business abstractions like dimensions, measures and fact tables. The semantic layer enabled the business users to have better understanding of their data environment, shielding them from the complexity of ordinary SQL interfaces.

This quickly changed when Business Intelligence (BI) tools introduced their own semantic layers (like the Business Objects Universe, or Cognos Catalog), generating a “multi-semantic environment” while at the same time the concept of a “Data Lake” was being born. This resulted in a multi-platform environment that forced users to move data between data platforms to make it accessible to the business. In addition, BI tools started providing data access to business users through proprietary semantic layers, allowing them to create multiple interpretations of the data within the same organization. Not surprisingly, self-service quickly became data chaos.

In an effort to resolve the data chaos problem, organizations started to notice the benefits of the data lake, amongst them the ability to store data as-is (structured or unstructured) without the need to transform it or move it to make it accessible to the business users. Enterprise organizations chose to migrate or establish their big data lakes on cloud technologies to lower costs, maximize their profits and increase performance.

IT organizations that are attempting to deliver Business Intelligence (BI) on big data for their users, often need to support multiple user types and multiple BI and visualization tools. For example, financial planners may access data using Excel, marketing analysts might be extensive users of Tableau, and enterprise reporting functions could be satisfied using MicroStrategy. Because each of these BI front ends has slightly different consumption patterns and interfaces, the resulting "real-life" BI-on-Big Data stack at any given enterprise could look very complex. AtScale makes Business Intelligence work on Big Data. With AtScale, business users get interactive and multi-dimensional analysis capabilities using the BI tools in which they have already invested.

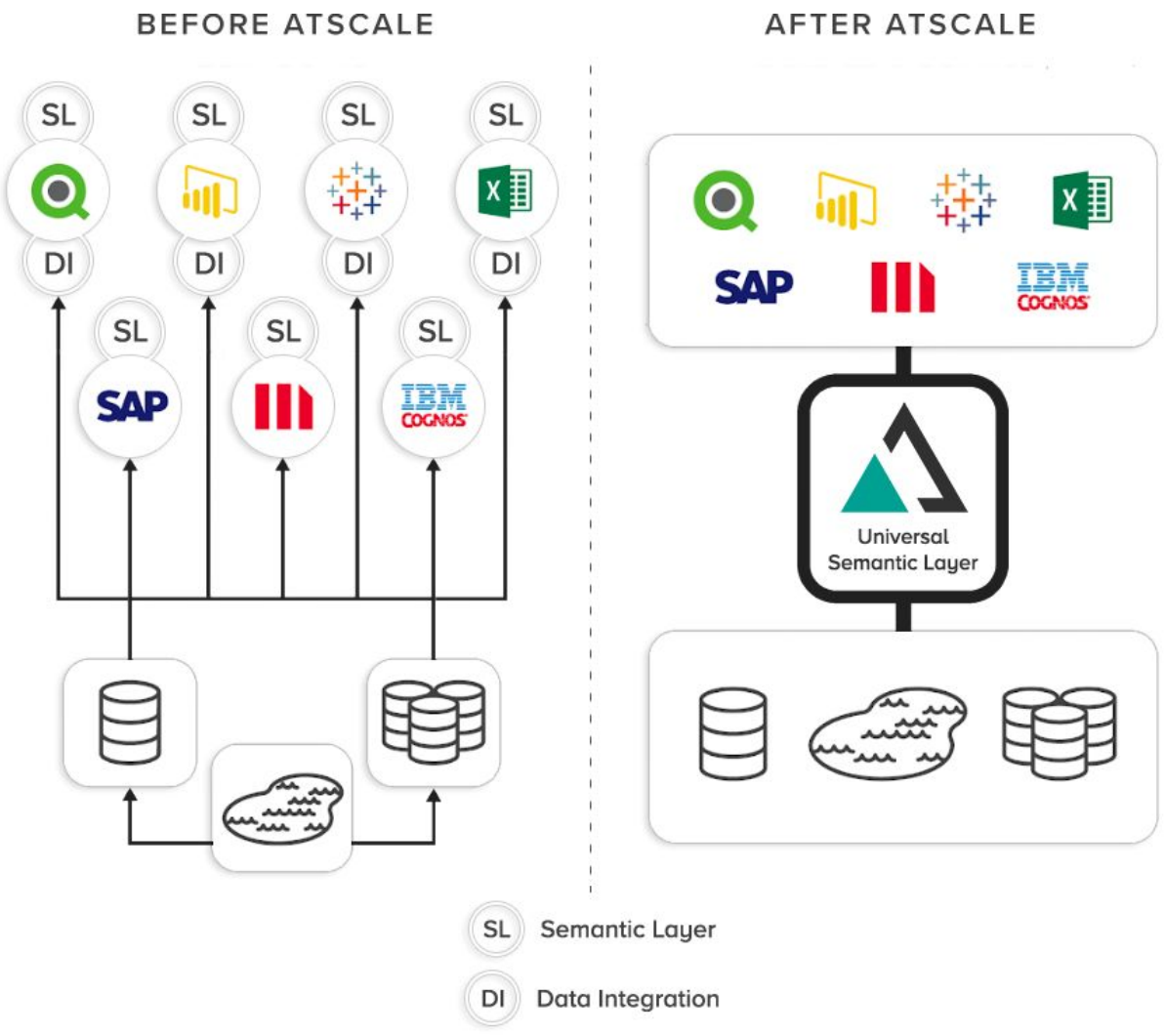


Figure 1. Evolution of The Modern Data Architecture.

Advantages of Google BigQuery in the Big Data World

In order to be successful, data driven organizations must analyze large data sets. This task requires computing capacity and resources that can vary in size depending on the kind of analysis or the amount of data being generated and consumed. This brings the challenge of installation and operational costs, as well as, the extra complexity of dynamically satisfying the demand for the extra resources needed to analyze variable amounts of data.

Cloud computing, with their common pay-per-use pricing model and ability to scale based on demand, makes it the most suitable candidate for this kind of big data workloads by easily delivering elastic compute and storage capability.

The ability to support analytics over petabyte-scale data, on a serverless enterprise data warehouse architecture makes Google BigQuery the cloud service of choice to deliver high-speed analysis on large data sets without requiring large onsite infrastructure investments, allowing BI users to cost-effectively address the most common data challenges.

Being part of the Google Cloud Platform (GCP,) Google BigQuery is designed to streamline big data analytics and storage, while removing the overhead cost and complexity of maintaining on-premises resources. Some of the advantages that Google BigQuery bring to the data world are:

- Cost Optimization
- Security
- Reliability
- Speed
- Scalability
- Simplicity

Cost Optimization

One of the things that makes Google BigQuery so attractive to big data consumers is its pay-as-you-go model approach allowing the user to pay for only the services used without requiring upfront costs or termination fees[1].

Google BigQuery is priced based on three different factors: 1) the amount of storage, 2) streaming inserts and 3) querying data, and it doesn't charge for loading and exporting data. In terms of storage, Google BigQuery pricing ranges anywhere from $0.02 per GB/month to $0.01 per GB/month for long term storage. Streaming inserts are priced at $0.01/200GB and queries are priced at $5/TB. A detailed breakdown of the pricing model is shown in figure 2.

US (multi-region) | Monthly

Operation	Pricing	Details
Active storage	$0.02 per GB	The first 10 GB is free each month. See Storage pricing for details.
Long-term storage	$0.01 per GB	The first 10 GB is free each month. See Storage pricing for details.
Streaming Inserts	$0.01 per 200 MB	You are charged for rows that are successfully inserted. Individual rows are calculated using a 1 KB minimum size. See Streaming pricing for details.
Queries (analysis)	$5 per TB	First 1 TB per month is free, see On-demand pricing for details. Flat-rate pricing is also available for high-volume customers.

Figure 2. Google BigQuery storage cost.

Security

Google Cloud Platform, including Google BigQuery, encrypts customer content stored at rest without any action required from the customer, using one or more encryption mechanisms. Data encryption is a process that takes readable data as input and transforms it into an output or code that reveals no context or information about the input and it is only readable by those who have access to its secret key.

For the Google Cloud Platform ecosystem, data is split into chunks, and each chunk is encrypted with a unique data encryption key which is stored with the data itself and can only be decoded using keys

[1] "Pricing - Google Cloud." https://cloud.google.com/bigquery/pricing.

managed by Google's central Key Management Service. This service is redundant and globally distributed.

Google considers security a high priority for GCP, its encryption at rest implementation reduces attacks and ensures that your data is securely stored and safe from unauthorized access[2].

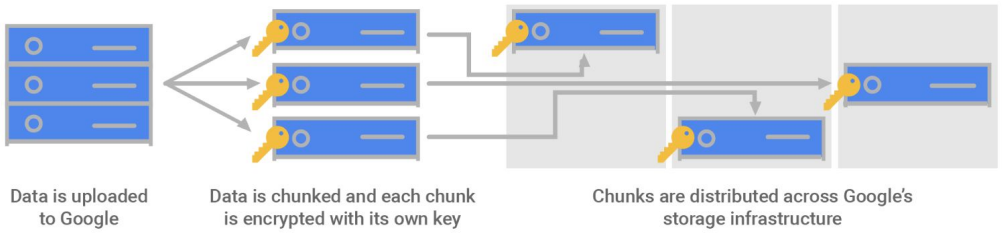


Figure 3. Google Cloud Platform Encryption Model.

Reliability

The Google Cloud Platform ensures an "always-on" availability and constant uptime with geo-replication across Google data centers. BigQuery has a dedicated global team of Site Reliability Engineers or SREs[3] who's job is to monitor the service for outages and performance failures. This team works behind the scenes to make sure that customers get the most of their current technology stack. If a failure or latency issue were to happen, SERs will seamlessly move data and processing queries to a different environment.

Speed

Google BigQuery allows to query, ingest and export petabyte - sized data sets with extraordinary speeds using GCP as the underlying cloud infrastructure. Naturally it performs on big data, being able to handle billions of rows without any errors in the order of 10s or 100's of seconds. The engine delivers interactive performance on known query pattern results in no more than several seconds on data sets sized in the order of millions of rows.[4]

[2] "Compliance - Google Cloud." https://cloud.google.com/security/compliance/.
[3] "Site Reliability Engineering - O'Reilly Media." http://shop.oreilly.com/product/0636920041528.do.
[4] "BI Performance Benchmarks with BigQuery from Google - AtScale Blog." 6 Apr. 2017, http://blog.atscale.com/bi-benchmarks-with-google-bigquery.

Scalability

GCP allows users to scale up to petabytes or down to kilobytes depending on the amount of data that users need to process, and requirements for performance and cost. BigQuery's elasticity allows it to scale to any size quickly and seamlessly, and as mentioned before, its cost model is based on the resources consumed.

Simplicity

Often, the most time consuming tasks for an analytics teams are data analysis and the data cleansing process. A step that many times gets overlooked, however, is setting up the environment. Being a fully managed data warehouse platform, GCP allows users to quickly complete all major tasks related to data analytics without the problem of having to manage the infrastructure. GCP provides ease of data loading, query tools, good out-of-the-box performance requiring minimal tuning or system configuration. These features accelerate time to value by allowing users to focus on analyzing their data rather than spending time configuring their systems for optimum performance.[5]

The Challenges

It is no surprise that many enterprise organizations are migrating to cloud technologies like the Google Cloud Platform because of the agility, scalability, security, ease of maintenance and cost savings provided by its services. However, organizations must consider whether the cloud platform can deliver on key functionality such as the ease of connecting BI tools, the ability to allow access to all of the data stored and the costs involved.

Four key challenges that organizations face with a move to the cloud are: cost, query concurrency, data modeling, and connectivity support. The ability to address these challenges can determine the success of a modern data strategy that includes the cloud.

[5] "BI Performance Benchmarks with BigQuery from Google - AtScale Blog." 6 Apr. 2017, http://blog.atscale.com/bi-benchmarks-with-google-bigquery.

Costs of Big Data Workloads

As more big data storage and processing workloads move to Google BigQuery, it is important to understand and manage the associated costs. Google BigQuery has a pay-as-you-go model which means that users can consume big data services without incurring large capital expenses. However, when the size of data starts to grow, the amount of resources utilized and cost grows as well.

There are three major elements to consider when estimating GBQ prices:

Storage Data

This is without a doubt one of the easiest to understand components of the GBQ pricing structure. Currently Google BigQuery charges a standard rate for all stored data of $0.02/GB/month. Consider the following scenario, if a user has 5TB of data (5000GB), the cost of storing it would look like this:

$$5000\ GB\ *\ \$0.02\ =\ \$100/month$$

Long Term Storage Data

It is a good practice to determine which data stored can be considered Long Term Storage (LTS). Google defines LTS as data that has not been updated in the last 90 consecutive days. Once data qualifies as a LTS, the price of storing it will drop by approximately 50 percent. Meaning the price will go from $0.02/GB/month to $0.01/GB/month.

Fundamentally a data set marked as a LTS is not any different than any other data set. This data set can be queried and have data exported from and its LTS status won't be affected. However, once a new record is added to it, the data set will be back on the original pricing bracket and the 90-day timer will be restarted.

Query Data Usage

The cost mentioned above ($0.02/GB or $0.01/GB for LTS) only covers storage, not queries. A user would have to pay separately per query based on the amount of data processed at a $5/TB rate.

For the sample data set being analyzed in this section, there are several parameters that need to be taken into consideration to estimate the query cost:

Parameter	Size/Cost
Cost per TB queried	$5
TB free per month	1
Fact table size	5TB
Number of columns in Fact	100
Number of users	100
Number of queries per day per user	10
Typical number of columns per query	5
Working days per user/month	20

Table 1. Google BigQuery cost estimation parameters.

In this scenario, once the parameters have been obtained, the cost estimation would look like this:

$$Fact\ Table\ Size * Number\ of\ users * Number\ of\ queries\ per\ day * Avg\ columns\ = TB\ queried\ per\ day$$

For simplicity of this analysis, the estimated number of users will be set at 100, running 10 queries per day, and an average of 20 working days per user. The estimation would look like follows:

$$5 * 100 * 10 * (5/100)\ = 250TB/Day$$

With a daily estimate of 250TB @ $5/TB, and factoring in the free TB that GBQ provides at time of writing, the processing cost on 20 working days plus storage cost would look like this:

$$Queried\ TB\ per\ day * Cost\ per\ TB\ queried * working\ days + storage\ cost\ = Total\ cost$$
$$250TB * \$5 * 20\ days + \$100\ = \$25,095$$

Operation	Pricing	Details
Active storage	$0.02 per GB	The first 10 GB is free each month. See Storage pricing for details.
Long-term storage	$0.01 per GB	The first 10 GB is free each month. See Storage pricing for details.
Streaming Inserts	$0.01 per 200 MB	You are charged for rows that are successfully inserted. Individual rows are calculated using a 1 KB minimum size. See Streaming pricing for details.
Queries (analysis)	$5 per TB	First 1 TB per month is free, see On-demand pricing for details. Flat-rate pricing is also available for high-volume customers.

Figure 4. Google BigQuery Pricing Model.

Query Concurrency

Highly available concurrency access is a must-have when implementing BI on the cloud and to deliver self-service BI to end users. According to Gartner, "Self-Service Analytics is a form of business intelligence (BI) in which line-of-business professionals are enabled and encouraged to perform queries and generate reports on their own, with nominal IT support."[ref4] Data driven organizations will allow users to generate as many queries as needed against their data lake, allowing many departments and teams trying to gain insights from the data they own.

Currently, Google BigQuery enforces a query concurrency limit of 50 queries, and a limit of 6 concurrent legacy SQL queries that contain user defined functions (UDF)[6].

Modeling

Data modeling is the key to success for BI on big data. Enabling end-users to navigate data without having to define a new query each time delivers consistent results in a timely manner. As well, when business users using disparate BI tools are given access to a centrally defined data model in a universal semantic layer, they can focus on the insights versus questioning the data.

Google BigQuery does not provide a user interface that allows the development of a data model to gain understanding of how data needs to flow. To ensure successful BI on big data strategies, a platform should deliver the ability to design a business-centered model easily allowing the user to access and obtain data from their data lake, define facts, dimensions and clear relationships between them to consume this data and gain insights through their BI tool of choice.

Connectivity Support

To perform data blending, visualizations and analytics, many analytics tools provide very powerful and flexible options. While the ultimate goal of each of these tools is to satisfy business needs, each tool has been created with a different focus. Because of this, each tool will use a different language to access data.

To deliver modern data architecture and analytics capabilities that allows a user to query the stored data, it is paramount that a cloud platform provide different connectivity options and protocols like ODBC, JDBC and APIs as well as support for query languages like SQL and MDX. Currently, Google

[6] "Quotas & Limits | BigQuery | Google Cloud." https://cloud.google.com/bigquery/quotas.

BigQuery does not natively support the MDX query language which is required for data mining purposes and provides great flexibility in different reporting services including Microsoft Excel.

Addressing BI Challenges with AtScale

Currently, businesses are finding more ways to turn data into value. In this drive to find value in data, companies are experiencing exponential growth of data. This growth can create data chaos. Data is getting larger because the storage cost is so low that it outweighs the potential gain that could be found in the data. However, according to Forrester, 75% of data remains unused. This is because companies are not equipped with the right architecture to give access to the data those who can find value in it.

According to the Ventana Research Cloud Analytics Benchmark[7], 48% of organizations already use cloud computing and 54% of data executives say that they have adopted cloud technologies with BI being the most important point of focus. In addition, today's average enterprise uses more than two BI tools[8] to ingest the data they have access to.

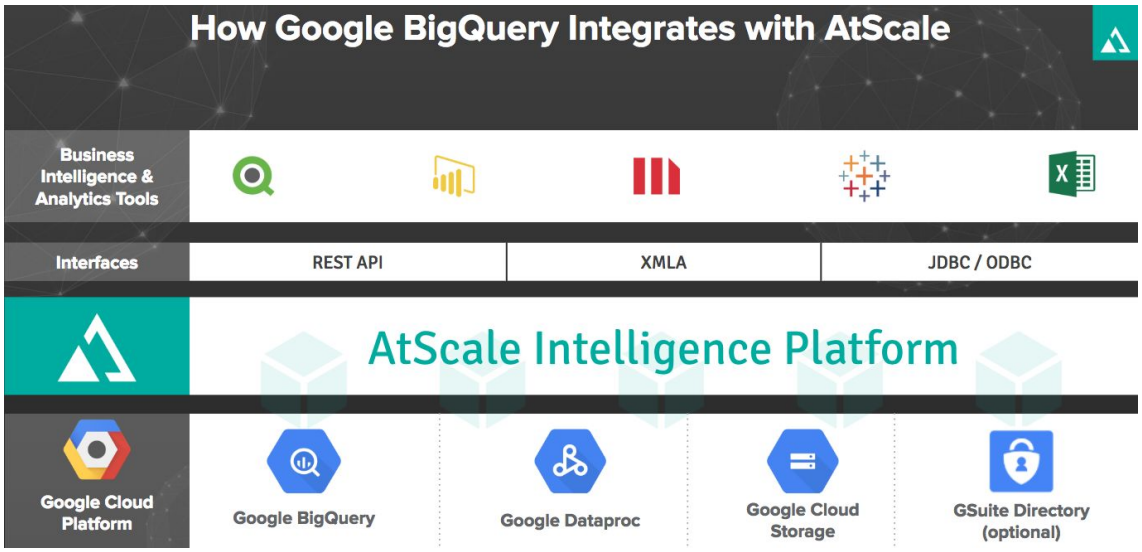


Figure 5. GBQ and AtScale integration.

While business users might have direct access to their organization's big data ecosystem through their BI tool of choice, each one of these tools is affected by the fundamental limitations of the cloud

[7] "Data and Analytics in the Cloud - Ventana Research." https://www.ventanaresearch.com/benchmark/big_data/data_and_analytics_in_the_cloud.
[8] "2018 Big Data Maturity Survey - AtScale." http://info.atscale.com/survey_2018_big_data_maturity_survey.

platform where the data resides. As a result, each tool uses a different engine to try to overcome these issues.

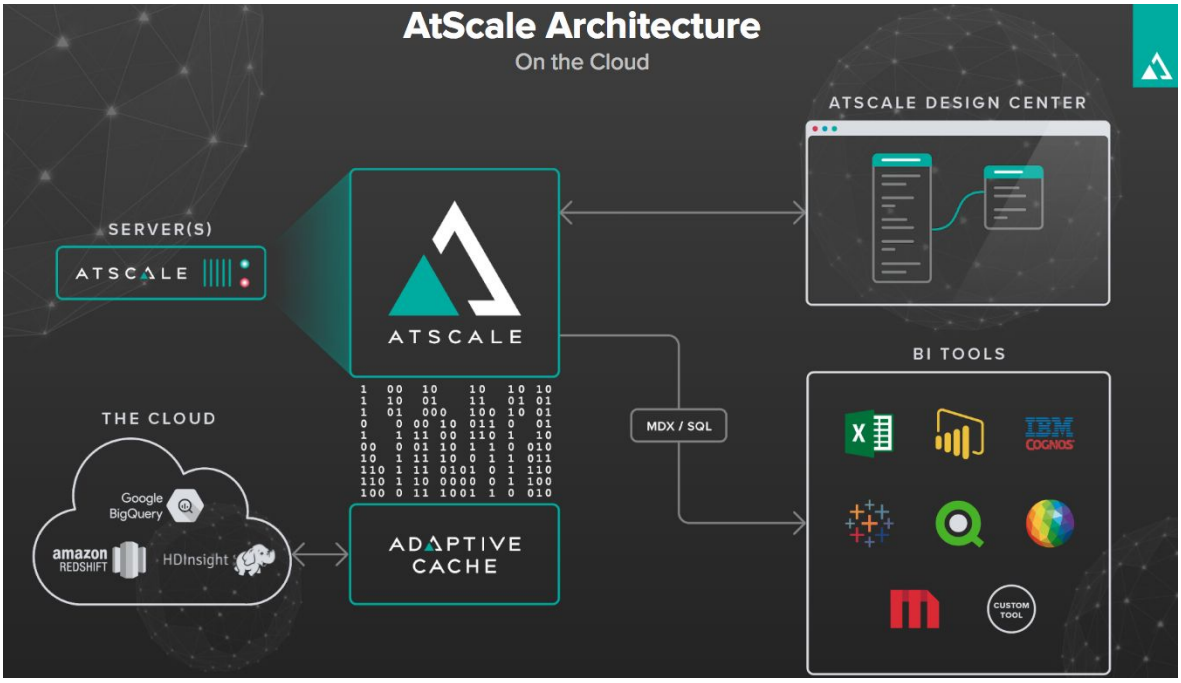


Figure 6. AtScale cloud architecture.

AtScale bridges the gap between the big data lake and the great variety of BI tools. Regardless of the type of access required (MDX, SQL, JDBC, ODBC or Rest API), AtScale provides a universal semantic layer for the big data environment. This provides context to the data, as well as consistency, fast query performance and data governance to all BI users on-premises and on the cloud.

Cost Management

One significant challenge when developing self-service big data analytics on the cloud is to deliver interactive performance for BI users while keeping costs low. AtScale incorporates a modern approach to solve this challenge by creating, managing, and optimizing aggregate tables.

These aggregate tables contain measures from one or more fact tables and include aggregated values for these measures. The aggregation of the data is at the level of one or more dimensional attributes, or, if no dimensional attributes are included, the aggregated data is a total of the values for the included measures.

AtScale aggregates reduce the number of rows that the query has to scan in order to obtain the results for the report or dashboard. By doing this, the length of time needed to produce the results

will be dramatically reduced. This results in reduced latency, as well as, a significant reduction in cloud resource consumption, translating into cost savings.

In the case of AtScale and Google BigQuery, to help manage costs, AtScale will make use of the aggregates stored on GBQ to produce results. Subsequent retrieval access is going to be very fast and with the cost of a very small footprint. Once the aggregate table has been built, subsequent access to raw data tables is avoided and the size of the Query Data Usage will be reduced.

To follow-up on the cost analysis shown in previous sections, the following calculations demonstrate how AtScale helps reducing the cost of data processing on GBQ. In order to provide a fair comparison, a extra number of parameters will be included in the calculation:

Parameter	Size/Cost (without AtScale)	Size/Cost (with AtScale)
Cost per TB queried	$5	$5
TB free per month	1	1
Fact table size	5TB	5TB
Number of columns in Fact	100	100
Number of users	100	100
Number of queries per day per user	10	10
Typical number of columns per query	5	5
Working days per user/month	20	20
Number of AtScale aggregates		100
Aggregate size (TB)		0.05
Minimum query size (TB)		0.05
Refresh Frequency of aggregates (in days)		1
TB Queries Per Day	250	50
TB Queried Per Month (user queries)	5000	1000
TB Queried Per Month (aggregate queries)		25
GBQ Monthly cost + storage	**$25,095**	**$5,220**

Table 2. Cost comparison of using GBQ with and without AtScale.

The savings after implementing AtScale are considerably large --this is because BI queries are no longer querying data directly from the raw fact tables. Instead, queries are being executed against AtScale aggregate tables which will be queried to obtain the same results at a fraction of the cost.

Business Modeling with AtScale

AtScale allows business users to design, develop and publish a multi-dimensional, relational model on top of the datasets stored in Google BigQuery. The AtScale virtual cube designer is based on the concepts that BI developers already understand. The AtScale model contains the metadata that BI applications need to browse and query data directly from GBQ. It makes the data lake look like any other multi-dimensional data mart or relational data warehouse, without the need for ETL, processing or moving the data out of the GCP environment.

According to Blue Hill Research[9] 40-60% of an analyst time is spent on data preparation. The AtScale universal semantic layer includes no data movement, minimizing the latency between IT and BI when preparing data for consumption, and reducing preparation time from weeks to minutes.

AtScale streamlines data delivery by eliminating repetitive movement for IT, automates traditionally manual processes, such as aggregate creation and maintenance and eliminates the cost related to labor and maintenance of the efforts involved with data movement and aggregation definitions.

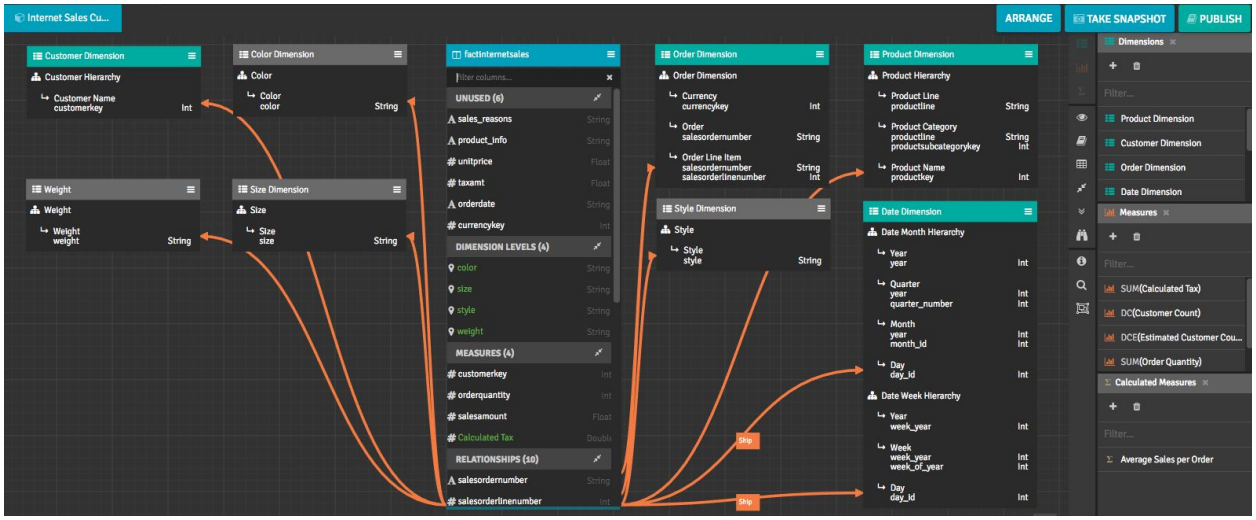


Figure 7. AtScale design canvas showing the complete Internet Sales virtual model.

Business users can design and publish these virtual models using the AtScale Design Center web application. Published models are immediately available to accept queries. Using standard

[9] "The Modern Data Analyst: Use Cases, Data Sources, and Time Spent ...." 3 Nov. 2015, http://bluehillresearch.com/the-modern-data-analyst-use-cases-data-sources-and-time/.

ODBC/JDBC drivers, users connect to a published cube in AtScale from existing BI tools like Tableau, PowerBI, Excel or Microstrategy and client applications. The AtScale engine intercepts SQL or MDX queries issued from BI tools and client applications, optimizes them, and executes them on Google BigQuery.

How is the concurrency issue addressed?

Empowering users with interactive modeling and data consumption frees up time to invest in the things that matter; complex analysis and deeper insights from data. However, the addition of many users who query data can translate into performance issues for the underlying data warehouse.

Every data warehouse possesses concurrency limitations -- the maximum number of queries that can be executed -- before the system reaches capacity. Concurrency issues can be experienced when trying to make data a shared asset among teams who are accessing the data to build reports or populate dashboards with their favorite BI tools. That, on its own, can be very taxing for the cloud platform.

Because the majority of queries generated by the AtScale Intelligence Platform will utilize aggregate tables, the cloud platform will handle queries that are hitting smaller data sets, thus allowing for faster execution, consequently allowing more queries per the amount of time, resulting in a larger throughput and increased query performance.

Enterprise big data analysis workflow

In the typical use case of consuming data from Google BigQuery using AtScale, data is ingested from its sources into the Google cloud storage. Users can utilize a variety of services in GCP to process, cleanse, de-normalize, classify, partition data. Once the data has been produced, it can be placed on Google BigQuery. AtScale will connect directly to GBQ not only to query data but also to place the aggregate tables that it will generate.

This process reduces the amount of data that users will query per day on GCP by more than 50% because AtScale uses aggregate definitions, which will be used to provide query results to the users without having to scan the entire raw dataset. Thereby, accelerating the process while reducing storage costs.

Figure 8. Enterprise big data analysis workflow

Use Case in Action

In this example, the Internet Sales sample data set consisting of five tables (one fact table, and 4 dimensions) will provide information about sales activity for a fictional goods website, where users can buy and sell any item ranging from books to articles of clothing. First, a virtual model was developed using the AtScale design canvas.

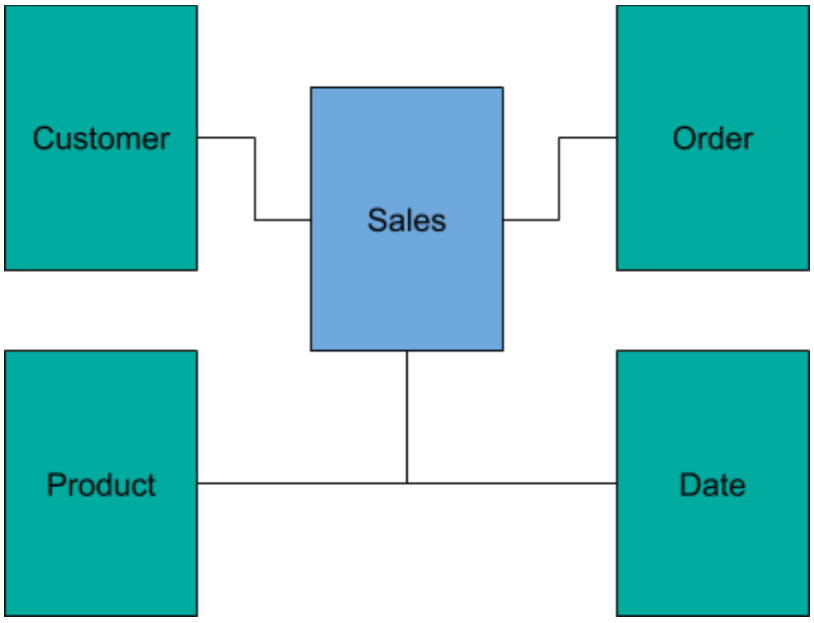


Figure 9. Internet Sales Sample Data Schema

As depicted in Fig. 9, a complete model will contain not only the fact table and dimensions that are needed to generate the report, but also the relationships between them, as well as measures and calculated fields that will be used to analyze the data in any of the supported BI tools.

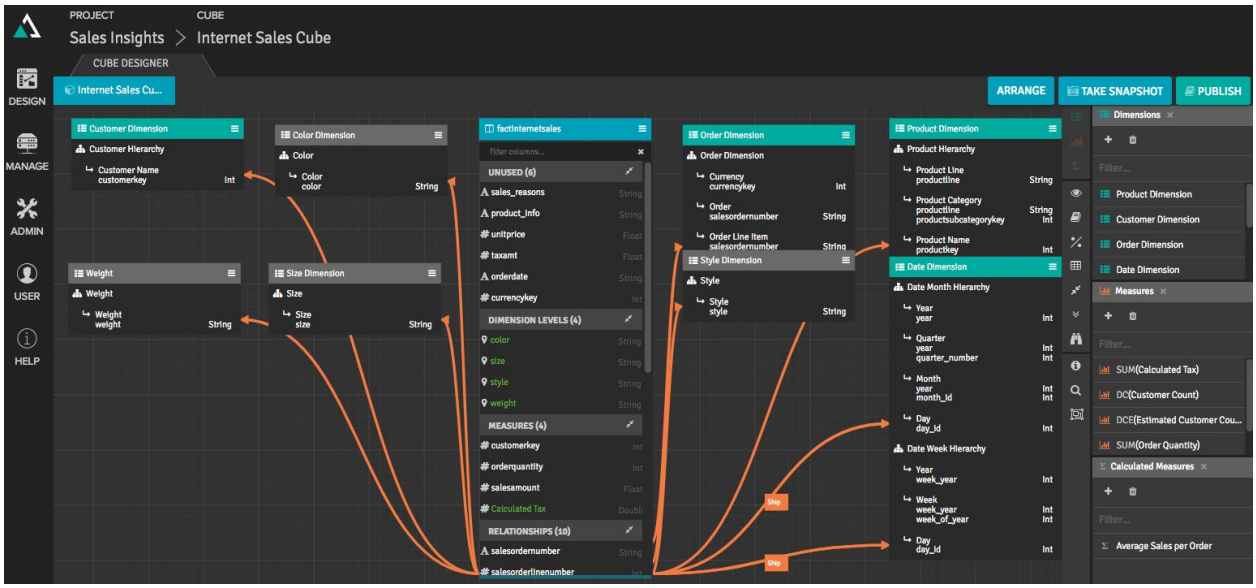


Figure 10. AtScale virtual model for the Internet Sales Sample data set.

Once the model is complete, the user can publish it, making it available to any supported BI tool. When AtScale publishes a model it generates a file that contains only the metadata for the underlying data source that the BI user will be working with. It does not contain any physical data and it is not a replication of the data that will be analyzed. By doing this, no data extracts are required and data stays securely in the cluster.

Queries from Microsoft Excel (Windows) and Tableau

Business intelligence users can start consuming data immediately using their BI tool of choice. Once the model has been completed and published from the AtScale Design Center, users will be presented with all the attributes and measures they need to analyze their data in their BI tool. These elements are the same (attributes and measures) that have been made available in the model created in the AtScale Design Center.

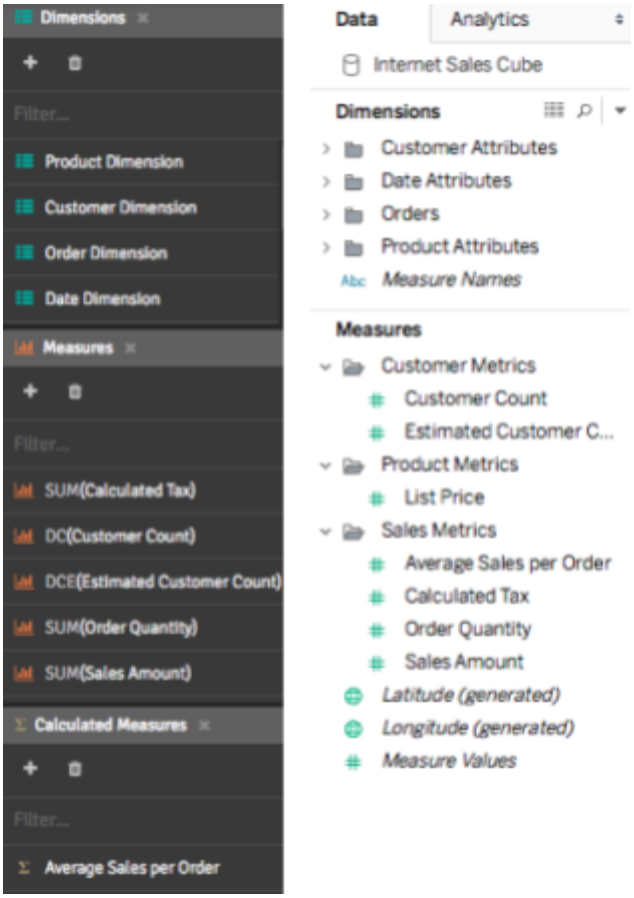


Figure 11. Comparison between cube elements created in AtScale vs those available in Tableau

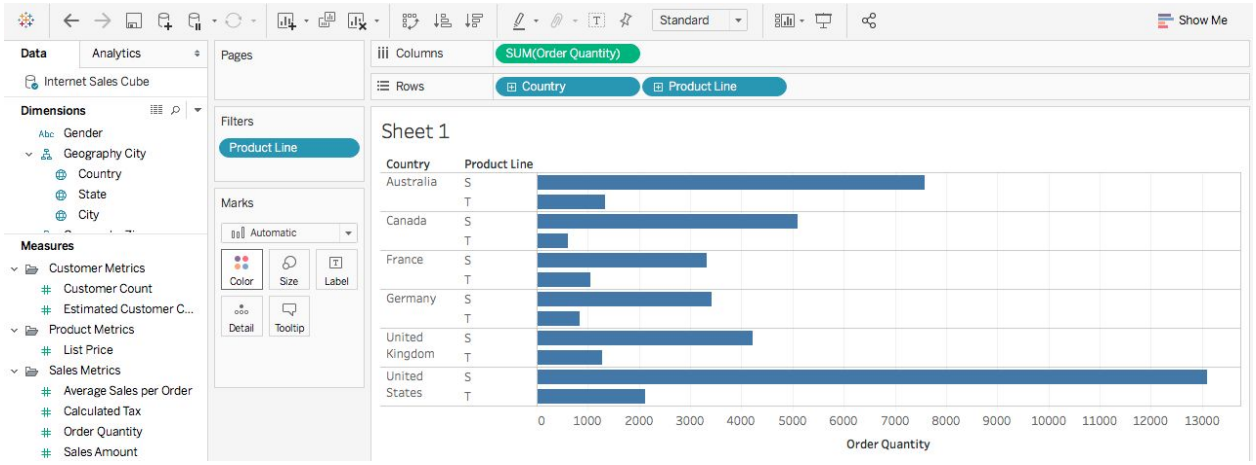


Figure 12. Tableau chart: Quantity of items per product line ordered by country.

A quick glance at the chart above allows the user to see the amount of per country. A filter has been applied to observe the data that relates to the product lines “S” and “T”.

The dynamics of the queries responsible for this report within the AtScale engine are shown below. Figure 13 shows the inbound query that AtScale is receiving from Tableau.

INBOUND QUERY

```
SELECT
  `Internet Sales Cube`.`CountryCity` AS `countrycity`,
  `Internet Sales Cube`.`Product Line` AS `product_line`,
  SUM(`Internet Sales Cube`.`orderquantity1`) AS `sum_orderquantity1_ok`
FROM
  `sales insights`.`internet sales cube` `Internet Sales Cube`
WHERE
  (
    `Internet Sales Cube`.`Product Line` IN ('', 'S ', 'T ')
  )
GROUP BY
  1,
  2
```

Figure 13. Inbound query from Tableau after applying a filter in the bar chart.

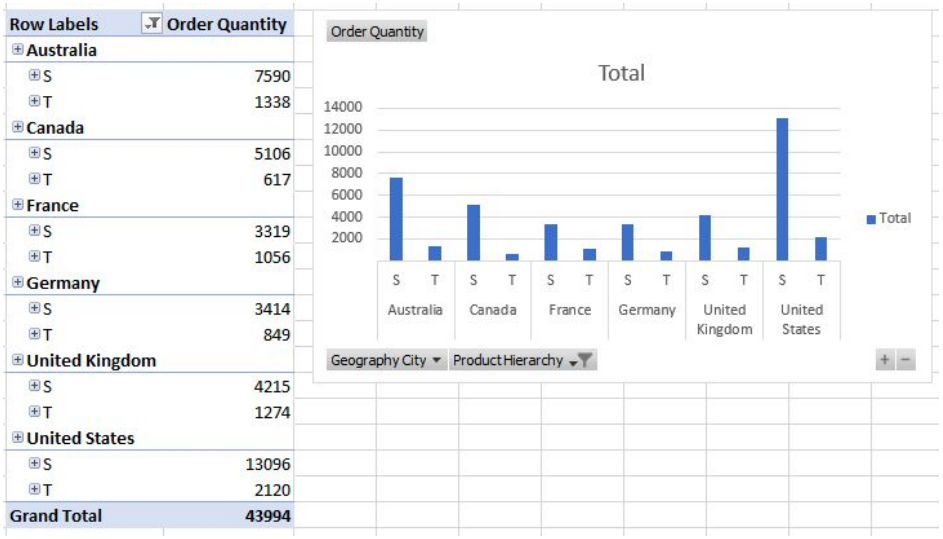
Figure 14 shows the outbound query that AtScale is using to obtain the data from Google BigQuery.

```
SELECT
  t_13.countrycity_topnc3 countrycity,
  t_13.product_line_topnc4 product_line,
  t_13.sum_orderquantity1_topnc5 sum_orderquantity1_ok
FROM
  (
    SELECT
      dim_geo_country_t12.country countrycity_topnc3,
      dimproduct_t7.productline product_line_topnc4,
      SUM(as_agg_28567add_no_t6.orderquantity1_c6) sum_orderquantity1_topnc5,
      dim_geo_country_t12.country countrycity_topnc1,
      dimproduct_t7.productline product_line_topnc2
    FROM
      as_adventure.as_agg_28567add_none as_agg_28567add_no_t6
      JOIN as_adventure.dimproduct dimproduct_t7 ON as_agg_28567add_no_t6.key_c1 = dimproduct_t7.productkey
      JOIN as_adventure.dimcustomer dimcustomer_t8 ON as_agg_28567add_no_t6.key_c2 = dimcustomer_t8.customerkey
      JOIN as_adventure.dim_geo_city dim_geo_city_t9 ON dimcustomer_t8.geographykey = dim_geo_city_t9.geographykey
      JOIN as_adventure.dim_geo_state dim_geo_state_t10 ON dim_geo_city_t9.statekey = dim_geo_state_t10.statekey
      JOIN as_adventure.dim_geo_postalcode dim_geo_postalcode_t11 ON dimcustomer_t8.geographykey = dim_geo_postalcode_t11.geographykey
      JOIN as_adventure.dim_geo_country dim_geo_country_t12 ON dim_geo_postalcode_t11.country = dim_geo_country_t12.country
    WHERE
      dimproduct_t7.productline IN ('', 'S ', 'T ')
    GROUP BY
      1,
      2,
      4,
      5
  ) t_13
```

Figure 14. Outbound query that AtScale will send to Google BigQuery to capture the filtered data.

One of the core principles that AtScale brings to business users is the ability to spend more time analyzing data and determining next steps, rather than making sure that the numbers yielded by different BI tools match. In this example, the same Internet Sales data is analyzed in Excel directly from GBQ.

When Excel connects to AtScale, Excel believes that it is connecting to a Microsoft Analysis Services Cube. In reality, it is accessing the same model that was used to create the previous report using



Row Labels	Order Quantity
Australia	
S	7590
T	1338
Canada	
S	5106
T	617
France	
S	3319
T	1056
Germany	
S	3414
T	849
United Kingdom	
S	4215
T	1274
United States	
S	13096
T	2120
Grand Total	43994

Figure 15. Pivot table generated in Excel (Windows).

Looking at the pivot table generated, the result for [fact and dimension] is the same that was yielded by Tableau. The queries are shown below.

INBOUND QUERY

```
SELECT
  NON EMPTY CrossJoin(
    Hierarchize(
      { DrilldownLevel(
        { [Geography Dimension].[Geography City].[All] },,,
        INCLUDE_CALC_MEMBERS
      ) }
    ),
    Hierarchize(
      { DrilldownLevel(
        { [Product Dimension].[Product Dimension].[All] },,,
        INCLUDE_CALC_MEMBERS
      ) }
    )
  ) DIMENSION PROPERTIES PARENT_UNIQUE_NAME,
  HIERARCHY_UNIQUE_NAME ON COLUMNS
FROM
  (
    SELECT
      (
        { [Product Dimension].[Product Dimension].[Product Line].[S ],
        [Product Dimension].[Product Dimension].[Product Line].[T ] }
      ) ON COLUMNS
    FROM
      [Internet Sales Cube]
  )
WHERE
  ([Measures].[orderquantity1]) CELL PROPERTIES VALUE,
  FORMAT_STRING,
  LANGUAGE,
  BACK_COLOR,
  FORE_COLOR,
  FONT_FLAGS
```

Figure 16. Inbound MDX query sent to AtScale from Excel (Windows)

Compared to the Tableau scenario, the inbound query in this case has been generated using MDX and then translated into the language supported by the underlying database.

```
SELECT
  t_7.country_c8 c0,
  dimproduct_t5.productline c1,
  SUM(as_agg_9edb8471_no_t4.orderquantity1_c6) c2
FROM
  as_adventure.as_agg_9edb8471_none as_agg_9edb8471_no_t4
  JOIN as_adventure.dimproduct dimproduct_t5 ON as_agg_9edb8471_no_t4.key_c2 = dimproduct_t5.productkey
  JOIN (
    SELECT
      dim_geo_country_t6.country country_c8
    FROM
      as_adventure.dim_geo_country dim_geo_country_t6
    GROUP BY
      1
  ) t_7 ON as_agg_9edb8471_no_t4.key_c1 = t_7.country_c8
WHERE
  dimproduct_t5.productline = 'S '
  OR dimproduct_t5.productline = 'T '
GROUP BY
  1,
  2
```

Figure 17. Outbound query set to GBQ from AtScale.

AtScale Adaptive Cache and Google BigQuery

The AtScale Adaptive Cache intelligently adapts to query patterns and data characteristics to deliver speed-of-thought analysis on billions of rows of data. AtScale incorporates the basic data warehousing concept of aggregate tables into its capabilities. This functionality enhances the performance of the two reports that were created as illustrated below.

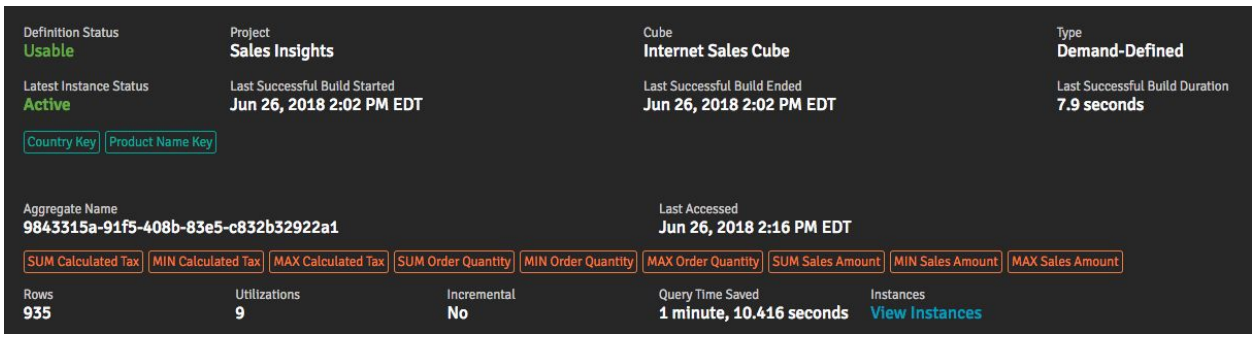


Figure 18. Outbound query sent to GBQ from AtScale.

AtScale generated the aggregation necessary to increase the performance of the queries that feeds the reports generated. These aggregate definitions will prevent the BI tool from scanning the raw data every time a similar query is generated.

Conclusion

As organizations become more data driven and more data is generated and stored on the cloud, BI professionals are challenged with the scenario of making sure that timely decisions can be made to have a positive outcome on the business. For this, it is imperative that the big data platform provides the robustness and flexibility necessary to face the hurdles presented.

In this document, we reviewed the elements that need to be taken into consideration when planning an implementation of BI on big data on the cloud as well as the challenges that such implementation can represent. Additionally, we discussed how AtScale helps improving the performance of the BI tool, productivity of the BI user while streamlining costs on Google BigQuery.

Furthermore, this paper analyzed use case scenario using the fictitious Internet Sales data set to observe how AtScale provides an enterprise-grade intelligence platform for any BI tool while enabling untethered data access and adhering to enterprise' performance, security and governance requirements.

About AtScale

AtScale is the industry leader in data federation and cloud transformation, enabling enterprises to modernize application architectures and accelerate business intelligence, A.I. and Machine Learning initiatives. By eliminating data location constraints, AtScale accelerates enterprise multi-platform and multi-cloud adoption with greater agility, performance and security -- all without disrupting the business. Led by industry veterans from Yahoo!, Google, Microsoft, Salesforce, Cisco and Oracle, AtScale is delivering enterprise transformation globally for firms including JPMorgan Chase, Wells Fargo, GlaxoSmithKline and many more. For additional information, visit www.atscale.com/cloud

AtScale Inc.
400 El Camino Real, Suite 800
San Mateo, CA 94402
info@atscale.com