ATSCALE

How to Maximize Your Business Value with AtScale and Amazon Redshift

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-structured, 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, among 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.

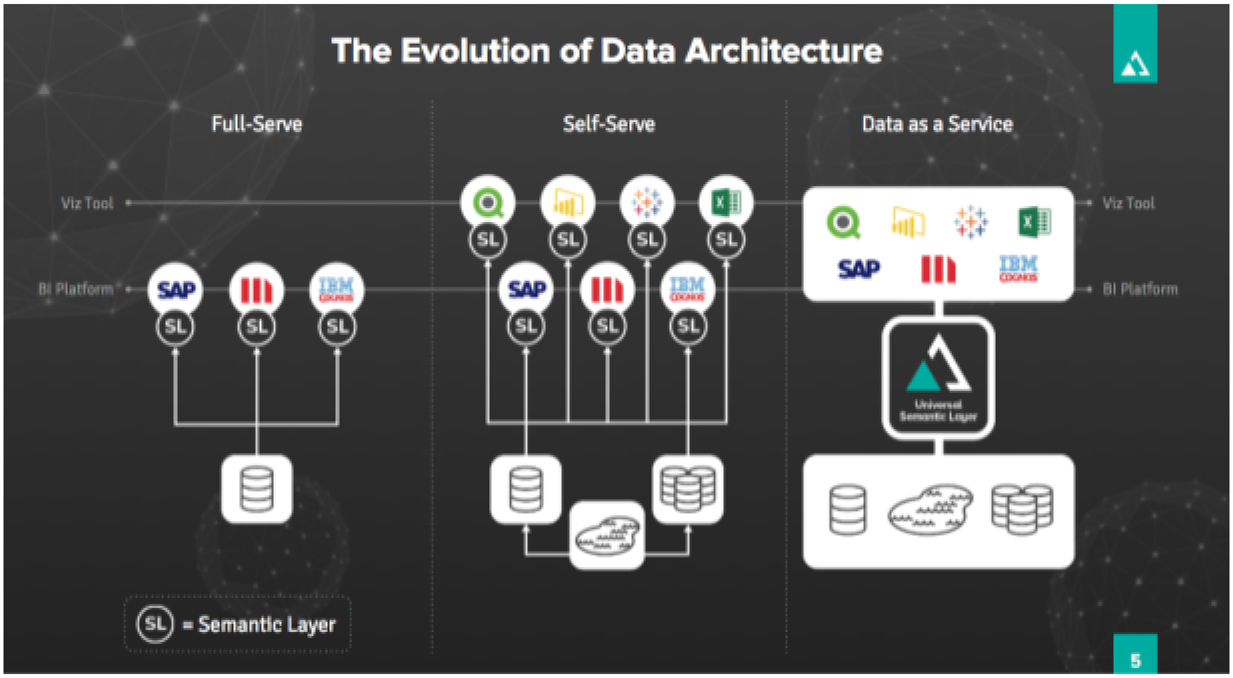


Figure 1. The Evolution of Data Architecture

Advantages of AWS in the Big Data World

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, and 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 expand and accommodate always increasing large-scale data volumes and its ability to process structured and unstructured data formats in the same environment, make AWS and its ecosystem of technologies a highly popular set of services designed to address common data challenges.

Data Collection and Storage

Amazon S3

Amazon Simple Storage Service (S3) has been designed with durability, scalability, and security as its main core principles. As a result, it is regularly used as a data storage technology for big data analytics, ranging from user-generated content to financial analysis. Amazon S3 provides a simple web services interface that can be used to easily store and retrieve any amount of data, at any time, from anywhere on the web. [[1]].

Among the many advantages that Amazon S3 provides, the most important ones are those that focus on the simplicity and robustness of the platform. It is designed to be a durable object store, it is designed to comply with the most exigent standards of availability and it can scale from petabytes to exabytes providing users with virtually unlimited storage capabilities. Additionally, for security and compliance, Amazon S3 users have many different ways to encrypt their objects. On the query side,

[1] "What Is Amazon S3? - Amazon Simple Storage Service - AWS ...." https://docs.aws.amazon.com/AmazonS3/latest/dev/Welcome.html.

Amazon S3 integrates with Amazon Redshift Spectrum, allowing users to retrieve data using SQL without the need to extract or load that data into another service.

Amazon Redshift

Amazon Redshift [2] is a fast, fully-managed, petabyte-scale data warehouse service that makes it simple and cost-effective to analyze data efficiently. Optimized for data sets ranging from several gigabytes to more than a petabyte, it delivers a cost-effective, elastic technology.

Amazon Redshift loads and distributes data into tables so that complex business queries can be executed efficiently. Amazon Redshift, through its Massive Parallel Processing (MPP) architecture and methodology, employs all available resources to ensure peak performance on SQL operations. Providing the most common connectivity options (ODBC, JDBC), Amazon Redshift is a versatile cloud-based data warehousing option.

On the query side, Amazon S3 integrates with Amazon Redshift Spectrum allowing users to retrieve data using SQL without the need to extract or load that data into another service. Amazon Spectrum gives users the ability to run fully SQL-supported Amazon Redshift queries against data stored in Amazon S3. It is fast, even at exabyte scale. It is highly available since it is a regional sub-service, is highly concurrent, and doesn’t require ETL since it allows for data to be queried in-place using open file formats without any sort of transformation.

Data Processing

Amazon Elastic MapReduce

Amazon Elastic MapReduce (EMR) provides a managed Hadoop framework that makes it easy, fast, and cost-effective to process vast amounts of data across dynamically scalable Amazon EC2 instances. [3] It can also run distributed frameworks such as Apache Spark and Presto, and interact with data in other AWS data stores such as Amazon S3.

Amazon EMR is commonly used to transform and clean the data from the original source format into the format needed at the destination. In this scenario, Hadoop provides the architecture to process the data and Amazon EMR provisions and manages the Hadoop cluster. Amazon EMR simplifies running Hadoop and related big data applications on AWS [4]. It eliminates the complexity and cost involved in managing a Hadoop installation. This creates the ability to spin up a Hadoop cluster in the

[2] "Amazon Redshift – Data Warehouse ...." https://aws.amazon.com/redshift/.
[3] "Amazon EMR – Amazon Web Services ...." https://aws.amazon.com/emr/.
[4] "Amazon EMR – Amazon Web Services ...." https://aws.amazon.com/emr/.

AWS cloud platform without having to sacrifice time and money on hardware provisioning. As a result, any user can start developing business analytics without incurring excessive costs.

Cost Model

One of the things that makes AWS so attractive to big data consumers is its pay-per-use model approach, allowing the user to pay for only the services used without requiring contracts or deposits.

Amazon Redshift pricing is based on the amount of data stored and the number of nodes in use. The number of nodes can be expanded or reduced depending on the amount of data users need to store or manage. Based on data volume stored, users can pick a setup anywhere from a single node, which is 160GB (0.016TB), to a 128 node cluster with a capacity of 16TB, ranging anywhere from $0.25/Hr to $6.80/Hr

Amazon S3 is priced based on three different factors: the amount of storage, monthly outbound data transfers, and the monthly number of data requests. The cost of storage is based on the total size of the objects in gigabytes stored in the Amazon S3 buckets, generally priced at $0.03/GB.

The Challenges

It is no surprise that many enterprise organizations are migrating to cloud platforms like Amazon AWS because of the agility, scalability, security, ease of maintenance, and cost savings provided by AWS. 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: query concurrency, data modeling, connectivity support, and costs. The ability to address these challenges can determine the success of a modern data strategy that includes the cloud.

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.”[[5]] Data driven organizations will

[5] "Self-Service Analytics - Gartner IT Glossary." https://www.gartner.com/it-glossary/self-service-analytics/.

allow users to generate as many queries as needed against their data lake, allowing many departments and teams to gain insights from the data they own.

Currently, Amazon Redshift enforces a query concurrency limit of 50 queries on a single cluster [6]. Queries are executed in a queue. By default there is one queue per query cluster which can run up to five concurrent queries. Users can only modify the configuration to allow up to 50 queries per queue and a maximum of 8 queues. However, this maximum query limit cannot be increased.

Modeling

Data modeling is the key to success for BI on big data. The ability to enable the end-user 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.

To ensure successful BI on big data strategies, a platform should deliver the ability to easily design a business-centered model, 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.

Amazon Redshift is a distributed relational database system capable of performing queries efficiently over petabytes of data. This is achieved by the combination of highly parallel processing, columnar design, and targeted data compression. Amazon Redshift does not provide a user interface that allows the development of a data model to gain understanding of how data needs to flow.

Connectivity Support

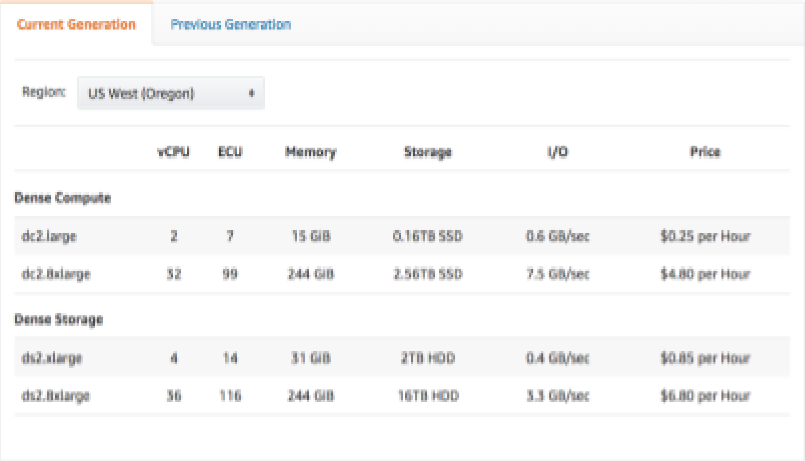
To perform data blending, visualizations, and analytics, many analytics tools provide very powerful and flexible options. While the ultimate goal of each of these tools is to satisfy business needs, each tool has been created with a different focus. Because of this, each tool will use a different language 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 provides different connectivity options and protocols like ODBC, JDBC, and APIs, as well as support for query languages like SQL and MDX. Currently, Amazon Redshift does not natively support the MDX query language, which is required for data mining purposes and provides great flexibility for different reporting services including Microsoft Excel.

[6] "Defining Query Queues - Amazon Redshift - AWS Documentation." https://docs.aws.amazon.com/redshift/latest/dg/cm-c-defining-query-queues.html.

Costs of Big Data Workloads

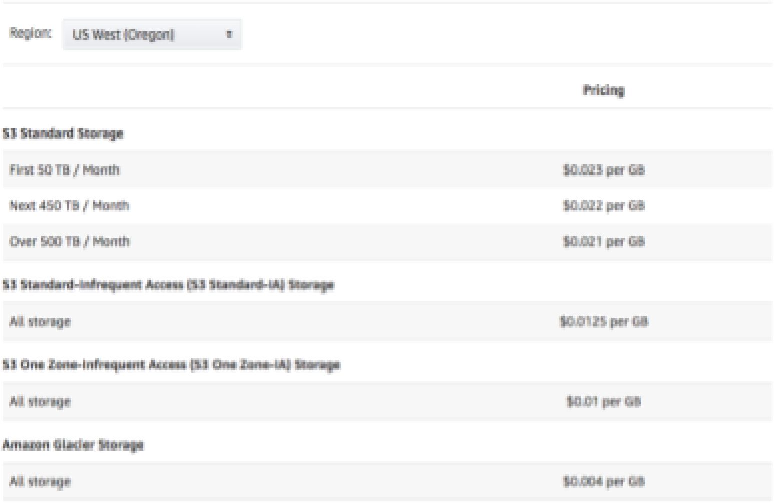
As more big data storage and processing workloads move to AWS, the enterprise must understand and manage the associated costs. Most AWS services have a pay-per-use model, which means that users can consume big data services without incurring large capital expenses. However, when the size of data starts to grow, the number of nodes and the price paid grows as well. The table below shows the pricing model of Amazon Redshift (fig. 2) compared to Amazon S3 (fig. 3). In a typical scenario, following the pricing model shown below, storing 2TB of data can cost $42/Month in Amazon S3 vs $620/Month on Amazon Redshift. As you can see, the most common data storage scenario can become expensive if a proper storage strategy is not put in place.



Region: US West (Oregon)

	vCPU	ECU	Memory	Storage	I/O	Price
Dense Compute						
dc2.large	2	7	15 GiB	0.16TB SSD	0.6 GB/sec	$0.25 per Hour
dc2.8xlarge	32	99	244 GiB	2.56TB SSD	7.5 GB/sec	$4.80 per Hour
Dense Storage						
ds2.xlarge	4	14	31 GiB	2TB HDD	0.4 GB/sec	$0.85 per Hour
ds2.8xlarge	36	116	244 GiB	16TB HDD	3.3 GB/sec	$6.80 per Hour

Figure 2. Amazon Redshift Pricing Model



Region: US West (Oregon)

	Pricing
S3 Standard Storage	
First 50 TB / Month	$0.023 per GB
Next 450 TB / Month	$0.022 per GB
Over 500 TB / Month	$0.021 per GB
S3 Standard-Infrequent Access (S3 Standard-IA) Storage	
All storage	$0.0125 per GB
S3 One Zone-Infrequent Access (S3 One Zone-IA) Storage	
All storage	$0.01 per GB
Amazon Glacier Storage	
All storage	$0.004 per GB

Figure 3. Amazon S3 Pricing Model

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.

[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 LP - AtScale." http://info.atscale.com/survey_2018_big_data_maturity_survey.

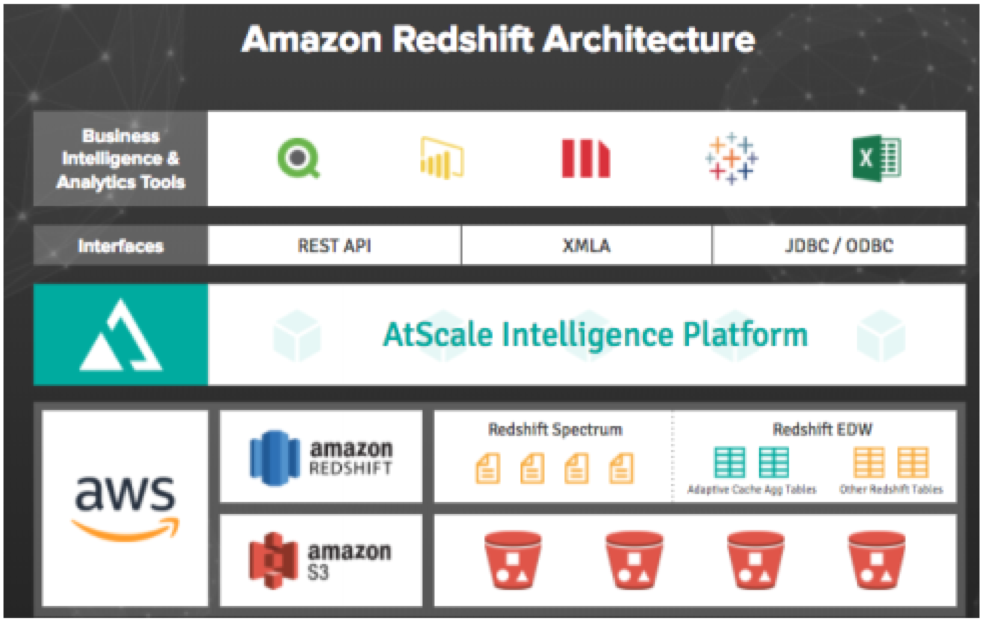

Amazon Redshift Architecture
Business Intelligence & Analytics Tools
Interfaces
REST API
XMLA
JDBC / ODBC
AtScale Intelligence Platform
aws
amazon REDSHIFT
amazon S3
Redshift Spectrum
Redshift EDW
Adaptive Cache Agg Tables
Other Redshift Tables

Figure 4. AtScale and Amazon Redshift Architecture Diagram

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

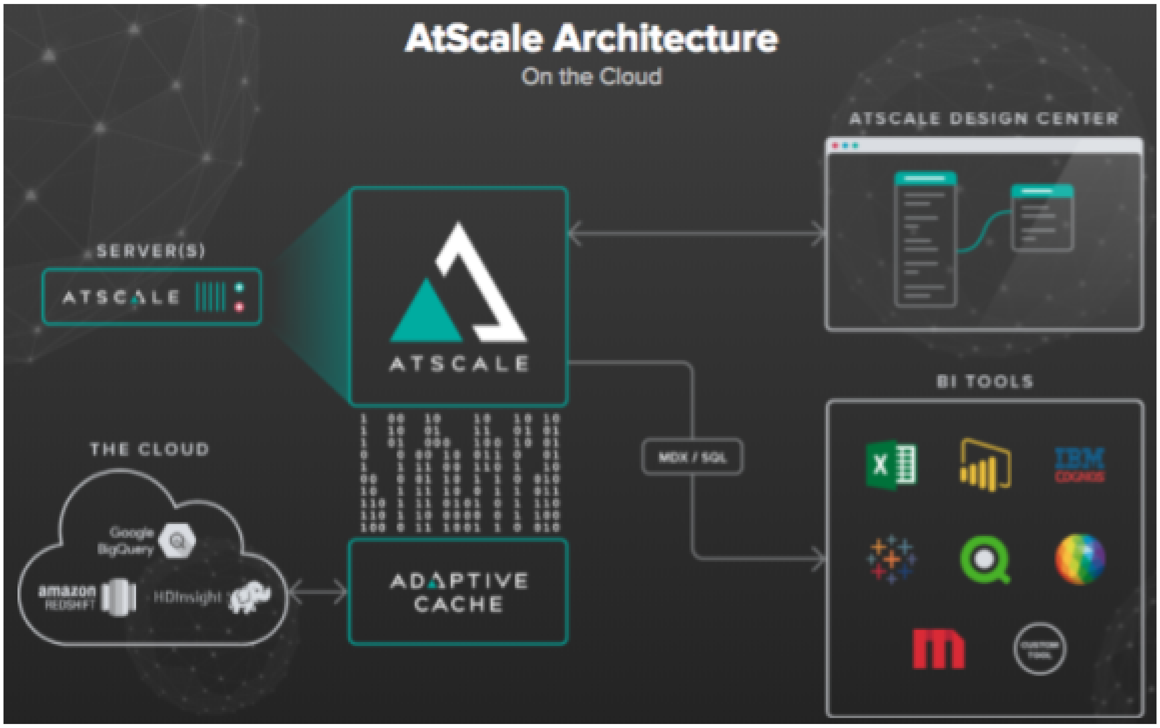


Figure 5. AtScale Cloud Architecture

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, consistency, unbelievably fast query performance, and data governance to all BI users on-premises and on the cloud.

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 Amazon Redshift. 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 Amazon Redshift. 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 Amazon AWS 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 labor and maintenance cost related to data movement and aggregation definitions.

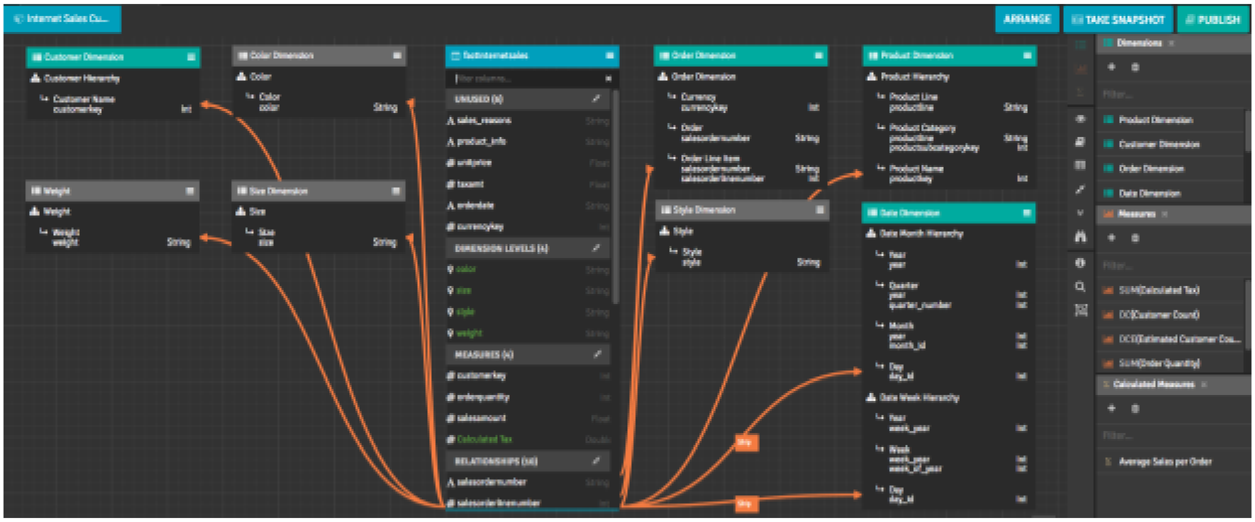


Figure 6. AtScale design canvas showing the complete Internet Sales virtual model.

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 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 Amazon Redshift by leveraging Amazon Spectrum.

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 on-cluster aggregate tables.

[9] "How Data Analysts Actually Spend Their Time (and Other Findings ...." 25 Aug. 2015, http://bluehillresearch.com/how-data-analysts-actually-spend-their-time-and-other-findings/.

These 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 and a significant reduction in cloud resource consumption, translating into increased savings.

For Amazon Redshift, to keep costs to a minimum, AtScale stores dimension tables in Amazon Redshift, while aggregates and fact tables are stored on Amazon S3 where storage costs are minimal.

Having dimensionality data on Amazon Redshift not only provides the benefits of a minimum footprint on AWS, but also avoids having to go back to Amazon Redshift to obtain results. AtScale makes use of the aggregates stored on Amazon S3 to produce results. Subsequent retrieval access is very fast and with the cost of a very small footprint. Once the aggregate table has been built, subsequent access to Amazon Redshift is avoided. This is an ideal model because it brings data movement to a minimum.

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. This delivers 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 Amazon S3 using AtScale, data is ingested from sources into Amazon S3. An Amazon EMR cluster is used to process, cleanse, de-normalize, classify, and partition data. Once that snapshot of data has been produced, it is placed on Amazon Redshift. AtScale will connect directly to Amazon S3, which will be reached via Amazon Redshift Spectrum. Then AtScale aggregate tables will be promoted to Amazon Redshift.

This process reduces the footprint on the AWS platform by more than 50% because AtScale is using Amazon Redshift to store aggregate definitions rather than S3. The benefit of having the dimension data on Amazon Redshift is that the use of Amazon S3 is minimized, thereby, accelerating the process while reducing storage costs.

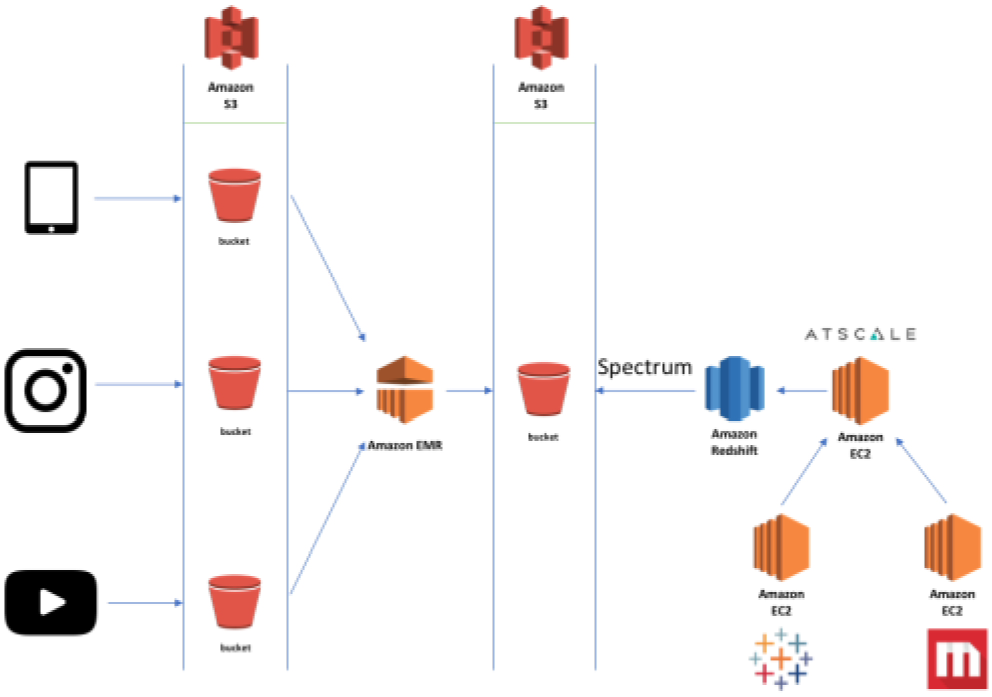


Figure 7. Enterprise big data analysis workflow.

Use Case in Action

In this example, the TICKIT sample data set[ref 8] consisting of seven tables (two fact tables, and five dimensions) will provide information about sales activity for the fictional TICKIT[[10]] website, where users buy and sell tickets online for sporting events, shows, and concerts. First, a virtual model was developed using the AtScale design canvas.

As depicted in Fig. 8, 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.

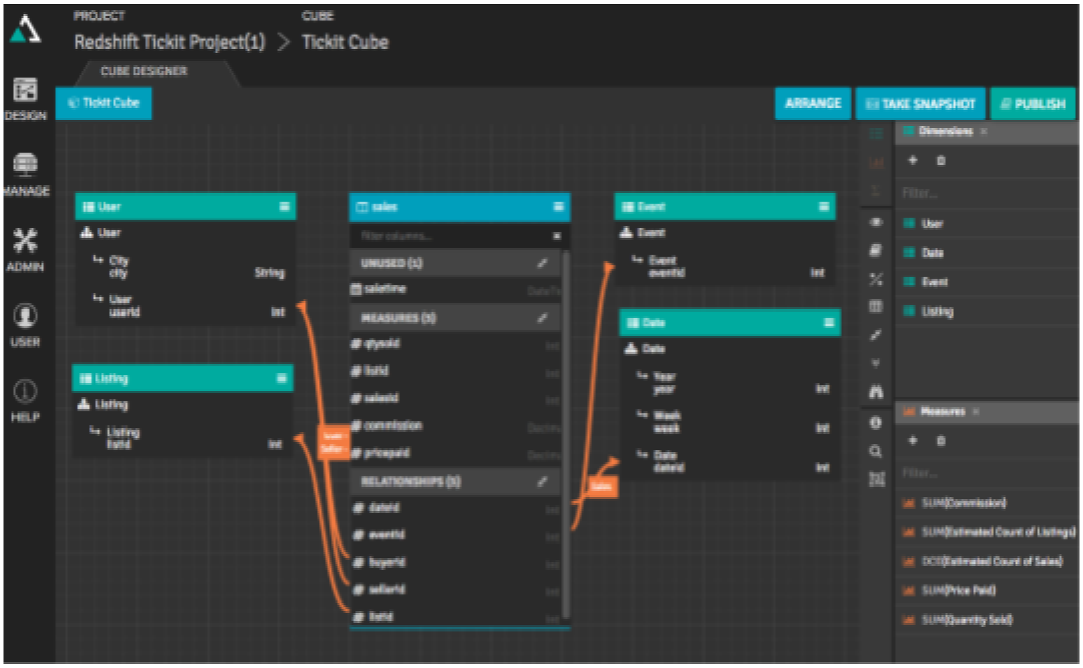


Figure 8. AtScale virtual model for the TICKIT data set.

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.

[10] "Sample Database - Amazon Redshift - AWS Documentation." https://docs.aws.amazon.com/redshift/latest/dg/c_sampledb.html.

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.[Fig. 9] These elements are the same (attributes and measures) that have been made available in the model created in the AtScale design center.

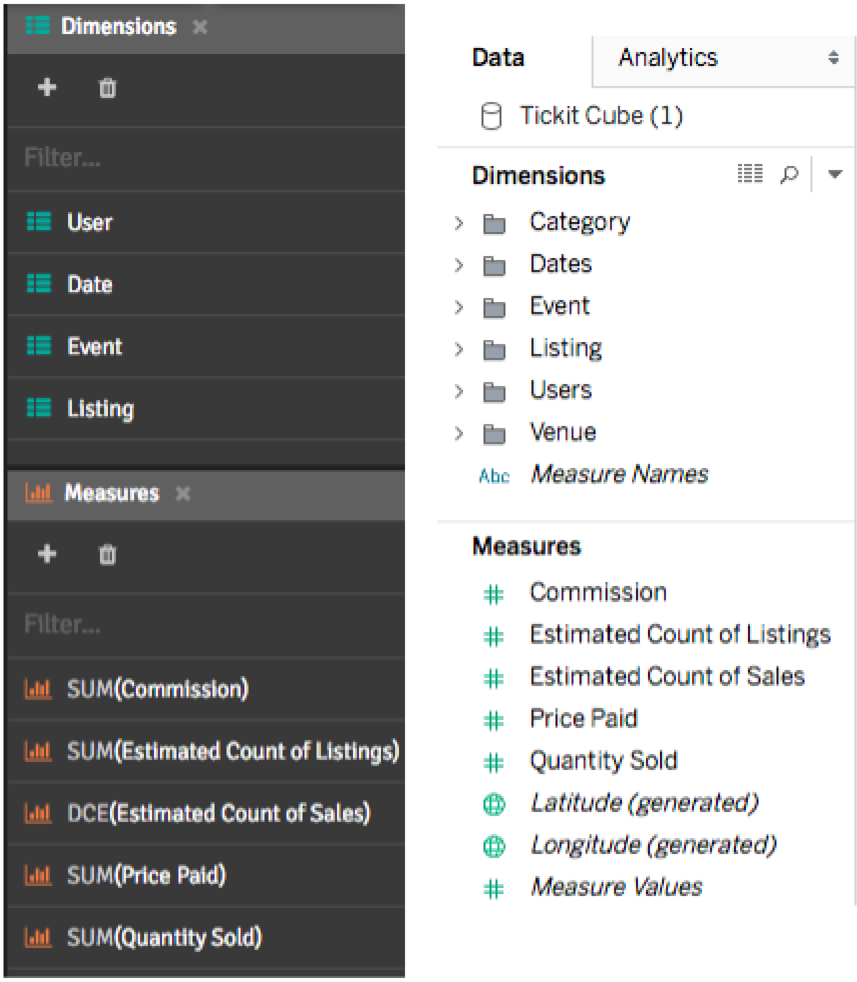


Figure 9. Comparison between cube elements created in AtScale vs those available in Tableau

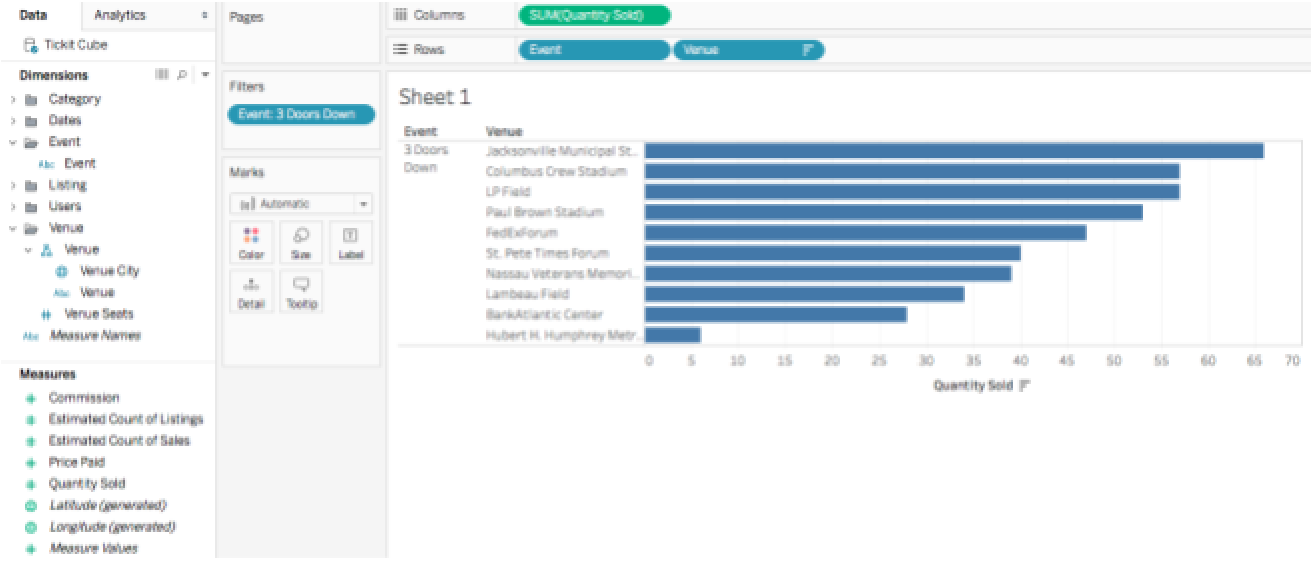
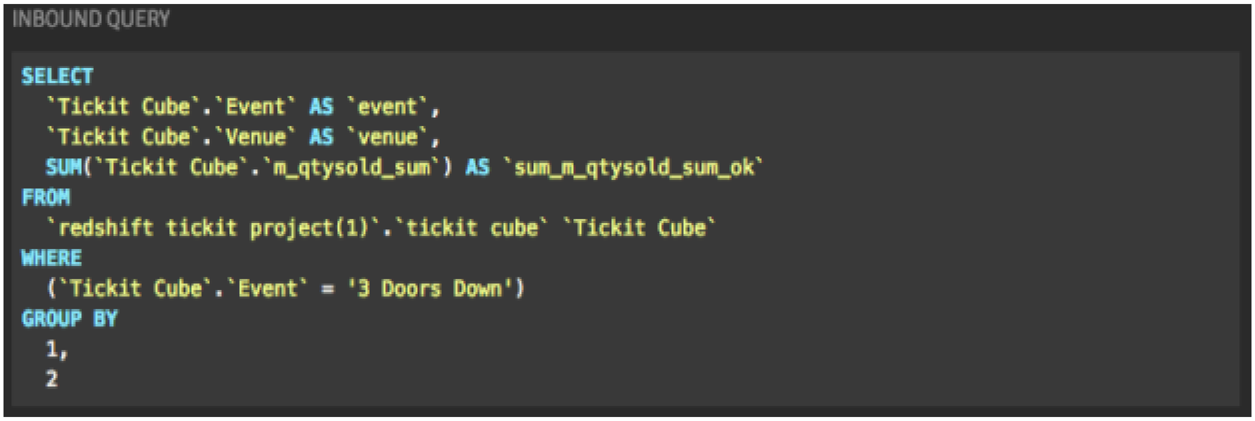


Figure 10. Tableau chart: Quantity of tickets Sold by Events per venue.

A quick glance at the chart above allows the user to see the amount of tickets sold per event and venue. A filter has been applied to observe the data that relates to the band “3 Doors Down”.

The dynamics of the queries responsible for this report within the AtScale engine are shown below. Figure 11 shows the inbound query that AtScale is receiving from Tableau.

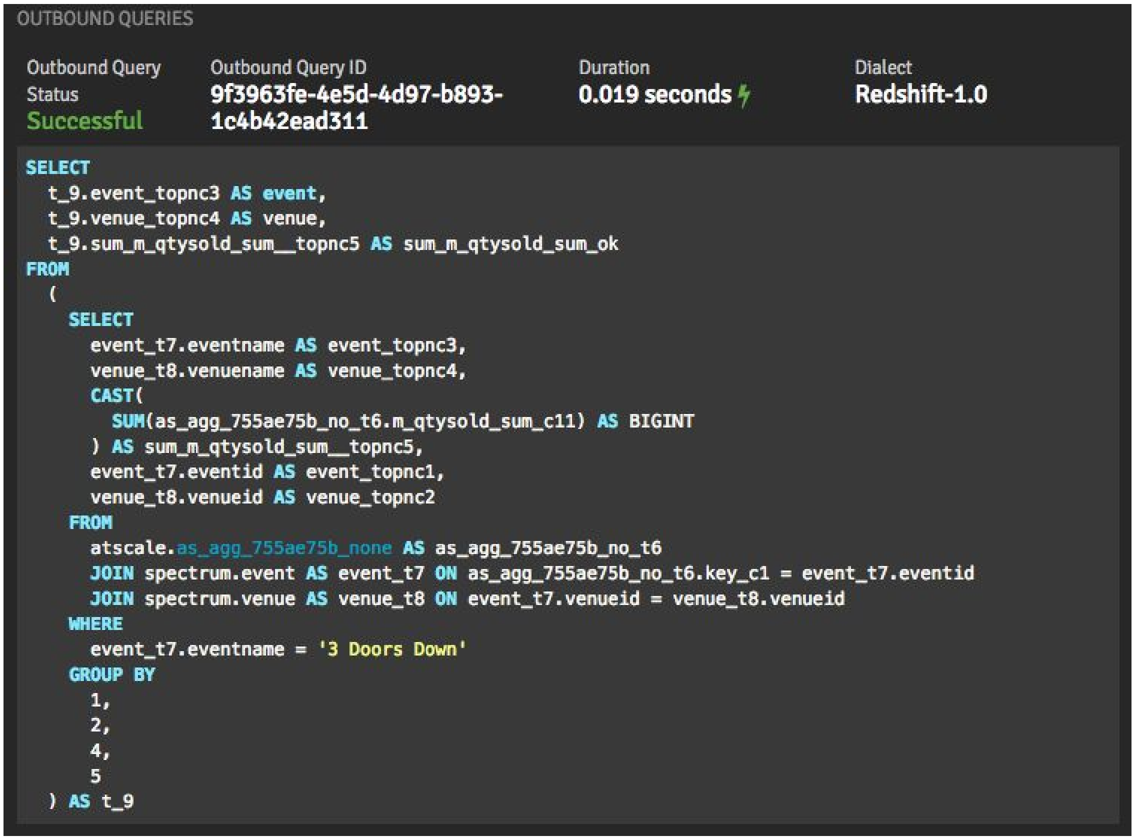


INBOUND QUERY

```
SELECT
  `Tickit Cube`.`Event` AS `event`,
  `Tickit Cube`.`Venue` AS `venue`,
  SUM(`Tickit Cube`.`m_qtysold_sum`) AS `sum_m_qtysold_sum_ok`
FROM
  `redshift tickit project(1)`.`tickit cube` `Tickit Cube`
WHERE
  (`Tickit Cube`.`Event` = '3 Doors Down')
GROUP BY
  1,
  2
```

Figure 11. Inbound query from Tableau after applying a filter in the bar chart.

Figure 12 shows the outbound query that AtScale is using to obtain the data using the Redshift dialect.


OUTBOUND QUERIES

Outbound Query Status
Successful

Outbound Query ID
9f3963fe-4e5d-4d97-b893-1c4b42ead311

Duration
0.019 seconds

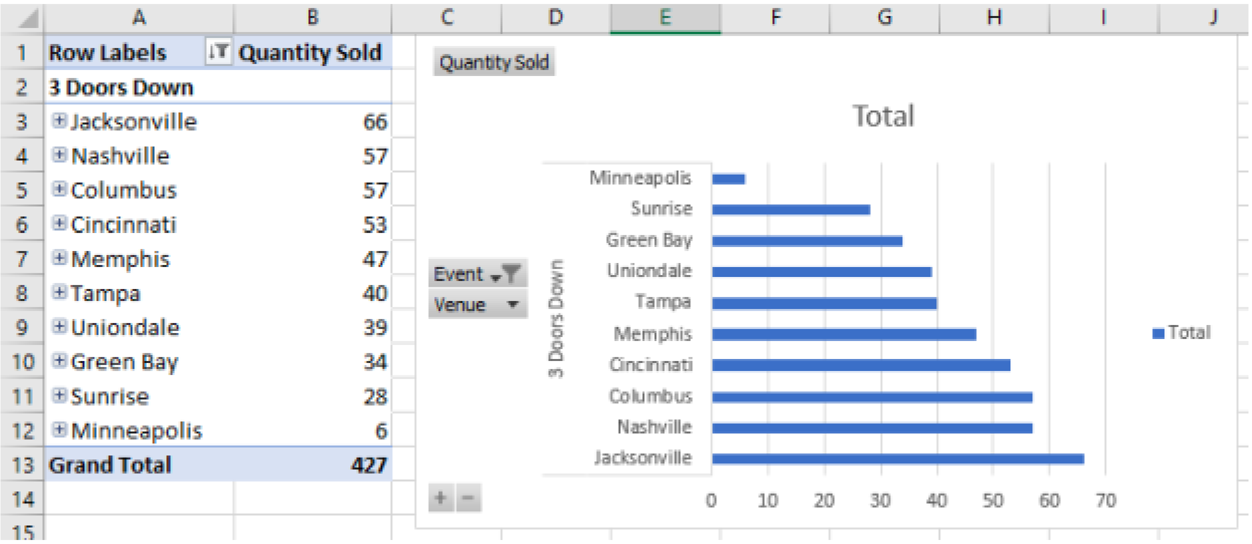
Dialect
Redshift-1.0

```
SELECT
  t_9.event_topnc3 AS event,
  t_9.venue_topnc4 AS venue,
  t_9.sum_m_qtysold_sum__topnc5 AS sum_m_qtysold_sum_ok
FROM
  (
    SELECT
      event_t7.eventname AS event_topnc3,
      venue_t8.venuename AS venue_topnc4,
      CAST(
        SUM(as_agg_755ae75b_no_t6.m_qtysold_sum_c11) AS BIGINT
      ) AS sum_m_qtysold_sum__topnc5,
      event_t7.eventid AS event_topnc1,
      venue_t8.venueid AS venue_topnc2
    FROM
      atscale.as_agg_755ae75b_none AS as_agg_755ae75b_no_t6
      JOIN spectrum.event AS event_t7 ON as_agg_755ae75b_no_t6.key_c1 = event_t7.eventid
      JOIN spectrum.venue AS venue_t8 ON event_t7.venueid = venue_t8.venueid
    WHERE
      event_t7.eventname = '3 Doors Down'
    GROUP BY
      1,
      2,
      4,
      5
  ) AS t_9
```

Figure 12. Outbound query that AtScale will send to Amazon Redshift 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 rather than making sure that the numbers yielded by different BI tools match. In this example, the same TICKIT data is analyzed in Excel directly from Redshift.

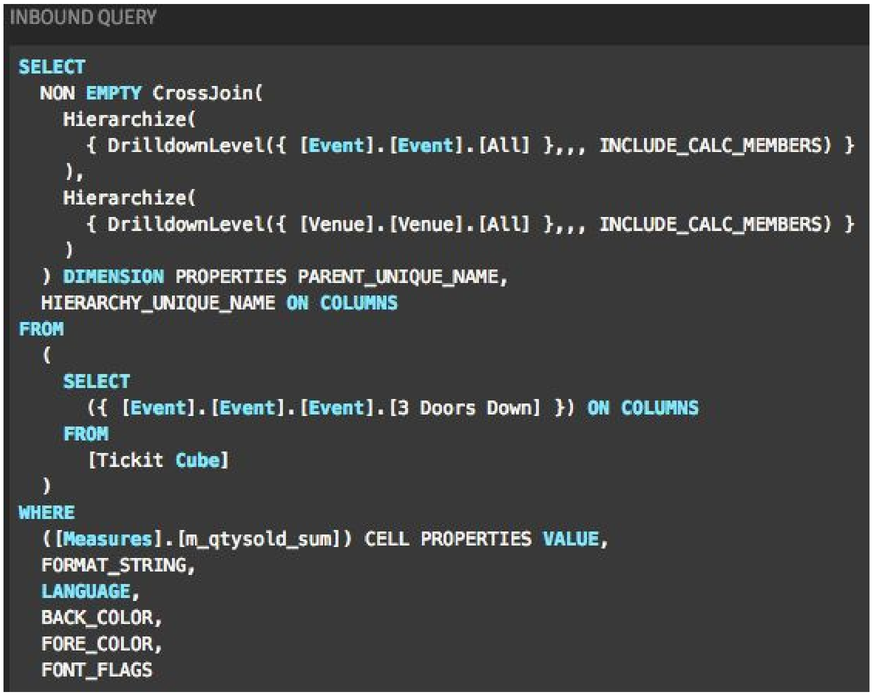
When Excel connects to AtScale, Excel believes that it is connecting to a Microsoft Analysis Services Cube. In reality, it is accessing the same model that was used to create the previous report using Tableau. In this case, the connection is done directly through JDBC and the queries are generated using MDX.



Row Labels	Quantity Sold
3 Doors Down	
⊞Jacksonville	66
⊞Nashville	57
⊞Columbus	57
⊞Cincinnati	53
⊞Memphis	47
⊞Tampa	40
⊞Uniondale	39
⊞Green Bay	34
⊞Sunrise	28
⊞Minneapolis	6
Grand Total	427

Figure 13. Tableau chart: Quantity of tickets Sold by Events per venue.

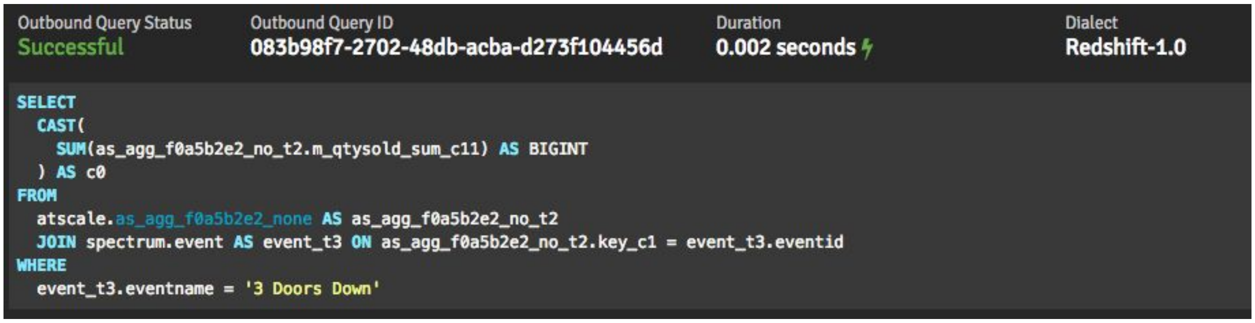
Looking at the live pivot table generated, the result for ticket sales per venue where the band is "3 Doors Down" is the same that was yielded by Tableau. The queries are shown below.



INBOUND QUERY

```
SELECT
  NON EMPTY CrossJoin(
    Hierarchize(
      { DrilldownLevel({ [Event].[Event].[All] },,, INCLUDE_CALC_MEMBERS) }
    ),
    Hierarchize(
      { DrilldownLevel({ [Venue].[Venue].[All] },,, INCLUDE_CALC_MEMBERS) }
    )
  ) DIMENSION PROPERTIES PARENT_UNIQUE_NAME,
  HIERARCHY_UNIQUE_NAME ON COLUMNS
FROM
  (
    SELECT
      ({ [Event].[Event].[Event].[3 Doors Down] }) ON COLUMNS
    FROM
      [Tickit Cube]
  )
WHERE
  ([Measures].[m_qtysold_sum]) CELL PROPERTIES VALUE,
  FORMAT_STRING,
  LANGUAGE,
  BACK_COLOR,
  FORE_COLOR,
  FONT_FLAGS
```

Figure 14. Inbound query that AtScale receives from Microsoft Excel in MDX format.



Outbound Query Status
Successful

Outbound Query ID
083b98f7-2702-48db-acba-d273f104456d

Duration
0.002 seconds

Dialect
Redshift-1.0

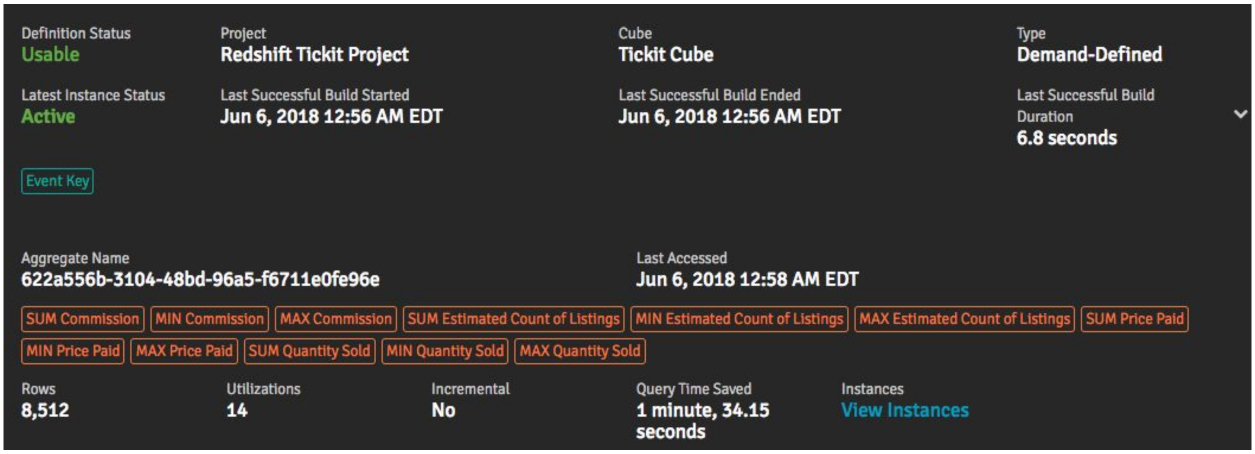
```
SELECT
  CAST(
    SUM(as_agg_f0a5b2e2_no_t2.m_qtysold_sum_c11) AS BIGINT
  ) AS c0
FROM
  atscale.as_agg_f0a5b2e2_none AS as_agg_f0a5b2e2_no_t2
  JOIN spectrum.event AS event_t3 ON as_agg_f0a5b2e2_no_t2.key_c1 = event_t3.eventid
WHERE
  event_t3.eventname = '3 Doors Down'
```

Figure 15. Outbound query that AtScale sends to the Amazon cluster in Redshift dialect.

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.

AtScale Adaptive Cache and Amazon AWS

AtScale incorporates the basic data warehousing concept of aggregate tables into its capabilities. This technology intelligently adapts to query patterns and data characteristics to deliver speed-of-thought analysis on billions of rows of data. This will prevent the BI tool from scanning the raw data every time a similar query is generated. As illustrated below this functionality enhances the performance of the two reports that were created. Figure 16 shows how the defined aggregate saved over 1 minute in query time. While it may not seem like much in a small example, the time and cost savings are exponentially magnified as the amount of data analyzed increases.



Definition Status
Usable

Project
Redshift Tickit Project

Cube
Tickit Cube

Type
Demand-Defined

Latest Instance Status
Active

Last Successful Build Started
Jun 6, 2018 12:56 AM EDT

Last Successful Build Ended
Jun 6, 2018 12:56 AM EDT

Last Successful Build Duration
6.8 seconds

Event Key

Aggregate Name
622a556b-3104-48bd-96a5-f6711e0fe96e

Last Accessed
Jun 6, 2018 12:58 AM EDT

SUM Commission | MIN Commission | MAX Commission | SUM Estimated Count of Listings | MIN Estimated Count of Listings | MAX Estimated Count of Listings | SUM Price Paid | MIN Price Paid | MAX Price Paid | SUM Quantity Sold | MIN Quantity Sold | MAX Quantity Sold

Rows
8,512

Utilizations
14

Incremental
No

Query Time Saved
1 minute, 34.15 seconds

Instances
View Instances

Figure 16. AtScale aggregate definition.

Conclusion

As organizations become more data driven and more data is generated and stored on the cloud, BI professionals are challenged to ensure 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 considered when planning an implementation of BI on big data on the cloud, and the challenges that such an implementation can represent. Additionally, we discussed how AtScale helps improving the performance of the BI tool and productivity of the BI user while streamlining costs on Amazon Redshift.

Furthermore, this paper analyzed a use case scenario using the fictitious TICKIT data set. In this analysis we observed 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
650.204.4959
info@atscale.com