



# **Amazon Web Services Data Engineering Immersion Day**

---

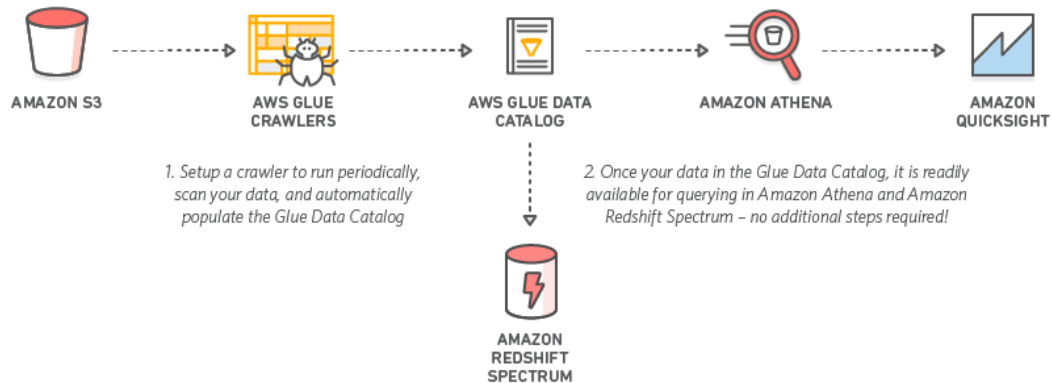
Exploring Data Lake with Amazon Athena and Amazon  
Quicksight  
*Jun 2019*

## Table of Contents

<b><i>Introduction.....</i></b>	<b><i>2</i></b>
Prerequisites.....	2
Getting Started.....	2
<b><i>Query Data with Amazon Athena .....</i></b>	<b><i>3</i></b>
<b><i>Build an Amazon QuickSight Dashboard.....</i></b>	<b><i>8</i></b>
Set up QuickSight.....	8
Create QuickSight Charts.....	11
Create QuickSight Parameters.....	13
Create a QuickSight Filter.....	15
Add Calculated Fields.....	17
<b><i>Amazon QuickSight ML-Insights (Optional).....</i></b>	<b><i>20</i></b>

## Introduction

This lab introduces you to AWS Glue, Amazon Athena, and Amazon QuickSight. AWS Glue is a fully managed data catalog and ETL service; Amazon Athena queries data; and Amazon QuickSight provides visualization of the data you import.



## Prerequisites

The DMS Lab and Glue ETL lab is a prerequisite for this lab.

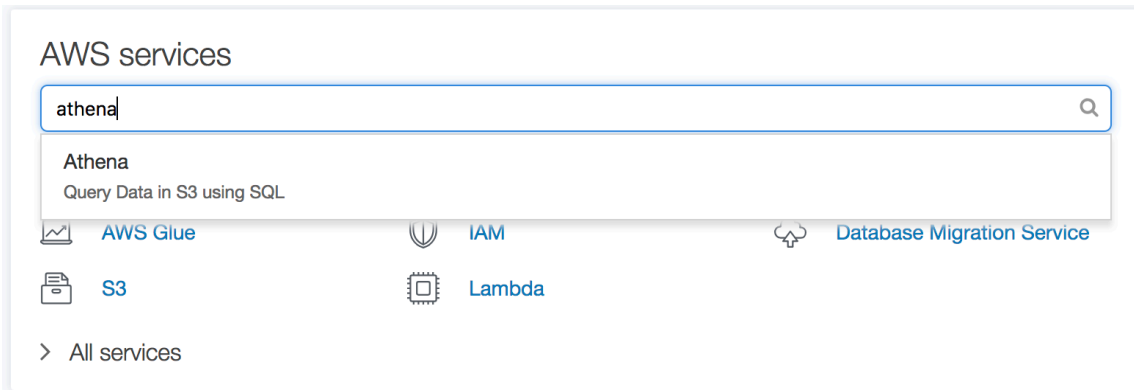
## Getting Started

In this lab, you will complete the following tasks:

1. [Query data and create a view with Amazon Athena](#)
2. [Build a dashboard with Amazon QuickSight](#)

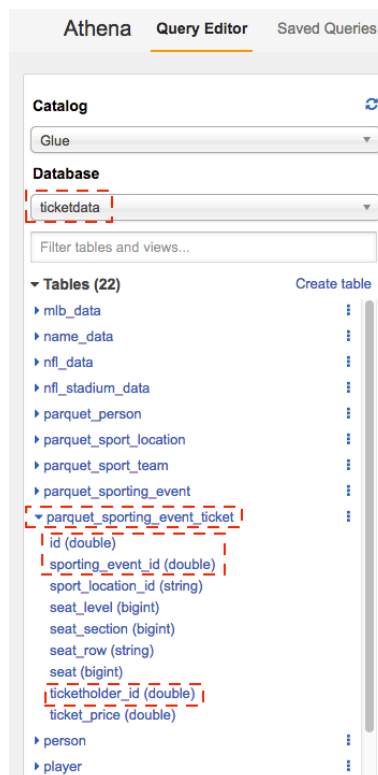
## Query Data with Amazon Athena

1. In the AWS services console, search for **Athena**.



2. In the Query Editor, select your newly created database e.g., "ticketdata".
3. Click the table named "parquet\_sporting\_event\_ticket" to inspect the fields.

**Note:** The type for fields *id*, *sporting\_event\_id* and *ticketholder\_id* should be (double).

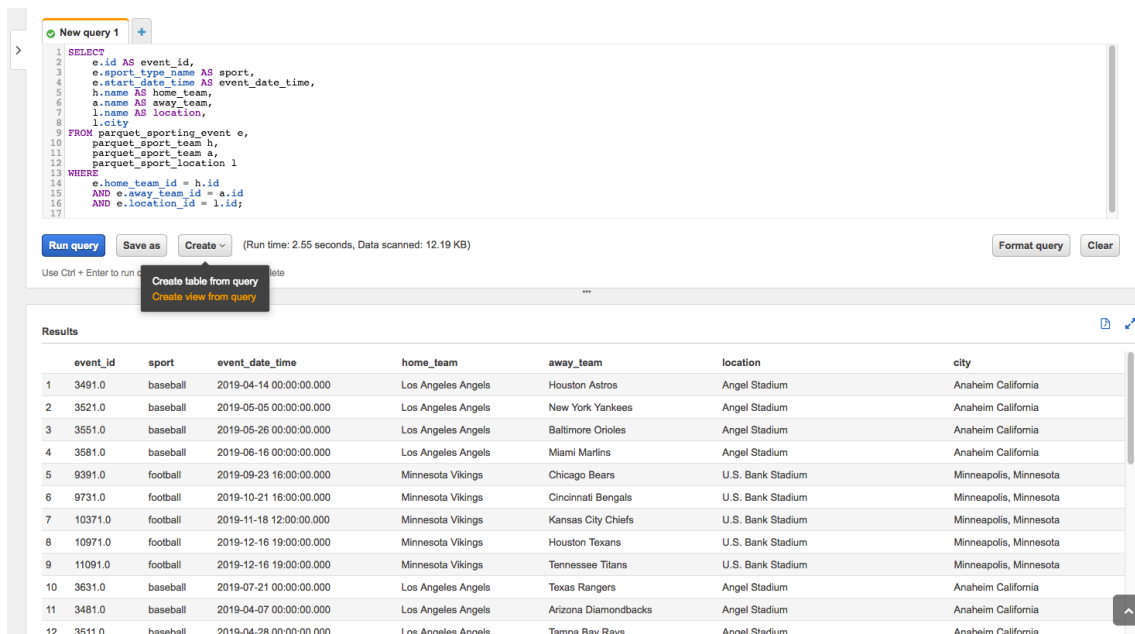


Next, we will query across tables `parquet_sporting_event`, `parquet_sport_team`, and `parquet_sport_location`.

4. Copy the following SQL syntax into the New Query 1 tab and click **Run Query**.

```
SELECT
    e.id AS event_id,
    e.sport_type_name AS sport,
    e.start_date_time AS event_date_time,
    h.name AS home_team,
    a.name AS away_team,
    l.name AS location,
    l.city
FROM parquet_sporting_event e,
    parquet_sport_team h,
    parquet_sport_team a,
    parquet_sport_location l
WHERE
    e.home_team_id = h.id
    AND e.away_team_id = a.id
    AND e.location_id = l.id;
```

The results appear beneath the query window.



The screenshot shows a SQL query editor interface. The query window displays the SQL code from the previous block. Below the query window, there are buttons for 'Run query', 'Save as', and 'Create'. A tooltip is visible over the 'Create' button, showing options: 'Create table from query' and 'Create view from query'. Below the query window, the 'Results' section displays a table with 12 rows of data. The table has columns: event\_id, sport, event\_date\_time, home\_team, away\_team, location, and city.

	event_id	sport	event_date_time	home_team	away_team	location	city
1	3491.0	baseball	2019-04-14 00:00:00.000	Los Angeles Angels	Houston Astros	Angel Stadium	Anaheim California
2	3521.0	baseball	2019-05-05 00:00:00.000	Los Angeles Angels	New York Yankees	Angel Stadium	Anaheim California
3	3551.0	baseball	2019-05-26 00:00:00.000	Los Angeles Angels	Baltimore Orioles	Angel Stadium	Anaheim California
4	3581.0	baseball	2019-06-16 00:00:00.000	Los Angeles Angels	Miami Marlins	Angel Stadium	Anaheim California
5	9391.0	football	2019-09-23 16:00:00.000	Minnesota Vikings	Chicago Bears	U.S. Bank Stadium	Minneapolis, Minnesota
6	9731.0	football	2019-10-21 16:00:00.000	Minnesota Vikings	Cincinnati Bengals	U.S. Bank Stadium	Minneapolis, Minnesota
7	10371.0	football	2019-11-18 12:00:00.000	Minnesota Vikings	Kansas City Chiefs	U.S. Bank Stadium	Minneapolis, Minnesota
8	10971.0	football	2019-12-16 19:00:00.000	Minnesota Vikings	Houston Texans	U.S. Bank Stadium	Minneapolis, Minnesota
9	11091.0	football	2019-12-16 19:00:00.000	Minnesota Vikings	Tennessee Titans	U.S. Bank Stadium	Minneapolis, Minnesota
10	3631.0	baseball	2019-07-21 00:00:00.000	Los Angeles Angels	Texas Rangers	Angel Stadium	Anaheim California
11	3481.0	baseball	2019-04-07 00:00:00.000	Los Angeles Angels	Arizona Diamondbacks	Angel Stadium	Anaheim California
12	3511.0	baseball	2019-04-28 00:00:00.000	Los Angeles Angels	Tampa Bay Rays	Angel Stadium	Anaheim California

5. As shown above Click **Create** and then select **Create view from query**

6. Name the view "sporting\_event\_info" and click **Create**.

## Create view

Views are updated each time you run a query

Name

Cancel Create

Your new view is created

Athena Query Editor

Saved Queries History AWS Glue Data Catalog Workgroup : primary

Catalog

Glue

Database

ticketdata

Filter tables and views...

Tables (22) Create table

Views (1) Create view

▼ sporting\_event\_info

event\_id (double)

sport (string)

event\_date\_time (timestamp)

home\_team (string)

away\_team (string)

location (string)

city (string)

New query 1

```

1 CREATE OR REPLACE VIEW "sporting_event_info" AS
2 SELECT
3     e.id AS event_id,
4     e.sport_type_name AS sport,
5     e.start_date_time AS event_date_time,
6     h.name AS home_team,
7     a.name AS away_team,
8     l.name AS location,
9     l.city
10 FROM parquet_sporting_event e,
11      parquet_sport_team h,
12      parquet_sport_team a,
13      parquet_sport_location l
14 WHERE
15     e.home_team_id = h.id
16     AND e.away_team_id = a.id
17     AND e.location_id = l.id;
18

```

Run query Save as Create (Run time: 1.16 seconds, Data scanned: 0 KB)

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

7. Copy the following SQL syntax into the New Query 2 tab and click **Run Query**.

```

SELECT t.id AS ticket_id,
       e.event_id,
       e.sport,
       e.event_date_time,
       e.home_team,
       e.away_team,
       e.location,
       e.city,
       t.seat_level,
       t.seat_section,
       t.seat_row,

```

```

t.seat,
t.ticket_price,
p.full_name AS ticketholder
FROM sporting_event_info e,
parquet_sporting_event_ticket t,
parquet_person p
WHERE
t.sporting_event_id = e.event_id
AND t.ticketholder_id = p.id

```

The results appear beneath the query window.

The screenshot shows the AWS Athena Query Editor interface. At the top, there's a navigation bar with 'Athena', 'Query Editor', 'Saved Queries', 'History', 'AWS Glue Data Catalog', and 'Workgroup: primary'. Below this, there's a toolbar with 'New query 1', 'New query 2', 'Run query', 'Save as', 'Create', 'Format query', and 'Clear'. The main area contains a SQL query that joins 'sporting\_event\_info', 'parquet\_sporting\_event\_ticket', and 'parquet\_person' tables. Below the query, there's a 'Results' section displaying a table with 14 columns: ticket\_id, event\_id, sport, event\_date\_time, home\_team, away\_team, location, city, seat\_level, seat\_section, seat\_row, seat, ticket\_price, and ticketholder. The table contains 12 rows of data for Philadelphia Eagles games.

ticket_id	event_id	sport	event_date_time	home_team	away_team	location	city	seat_level	seat_section	seat_row	seat	ticket_price	ticketholder
4320911.0	9881.0	football	2019-10-21 16:00:00.000	Philadelphia Eagles	Cincinnati Bengals	Lincoln Financial Field	Philadelphia, Pennsylvania	2	11	B	3	44.64	Corinne Buck
4320921.0	9881.0	football	2019-10-21 16:00:00.000	Philadelphia Eagles	Cincinnati Bengals	Lincoln Financial Field	Philadelphia, Pennsylvania	2	11	B	2	44.64	Corinne Buck
4320931.0	9881.0	football	2019-10-21 16:00:00.000	Philadelphia Eagles	Cincinnati Bengals	Lincoln Financial Field	Philadelphia, Pennsylvania	2	11	B	1	44.64	Corinne Buck
4330561.0	9881.0	football	2019-10-21 16:00:00.000	Philadelphia Eagles	Cincinnati Bengals	Lincoln Financial Field	Philadelphia, Pennsylvania	2	11	B	2	133.92	Corinne Buck
4317081.0	9881.0	football	2019-10-21 16:00:00.000	Philadelphia Eagles	Cincinnati Bengals	Lincoln Financial Field	Philadelphia, Pennsylvania	2	11	A	2	44.64	Corinne Buck
4317091.0	9881.0	football	2019-10-21 16:00:00.000	Philadelphia Eagles	Cincinnati Bengals	Lincoln Financial Field	Philadelphia, Pennsylvania	2	11	A	1	44.64	Corinne Buck
4320961.0	9881.0	football	2019-10-21 16:00:00.000	Philadelphia Eagles	Cincinnati Bengals	Lincoln Financial Field	Philadelphia, Pennsylvania	2	11	A	1	44.64	Corinne Buck
4320951.0	9881.0	football	2019-10-21 16:00:00.000	Philadelphia Eagles	Cincinnati Bengals	Lincoln Financial Field	Philadelphia, Pennsylvania	2	11	A	2	44.64	Corinne Buck
4320941.0	9881.0	football	2019-10-21 16:00:00.000	Philadelphia Eagles	Cincinnati Bengals	Lincoln Financial Field	Philadelphia, Pennsylvania	2	11	A	3	44.64	Corinne Buck
4320911.0	9881.0	football	2019-10-21 16:00:00.000	Philadelphia Eagles	Cincinnati Bengals	Lincoln Financial Field	Philadelphia, Pennsylvania	2	11	A	2	133.92	Corinne Buck
4330601.0	9881.0	football	2019-10-21 16:00:00.000	Philadelphia Eagles	Cincinnati Bengals	Lincoln Financial Field	Philadelphia, Pennsylvania	2	11	A	3	133.92	Corinne Buck
4330621.0	9881.0	football	2019-10-21 16:00:00.000	Philadelphia Eagles	Cincinnati Bengals	Lincoln Financial Field	Philadelphia, Pennsylvania	2	11	A	1	133.92	Corinne Buck

8. As shown above Click **Create view from query**.
9. Name the view "sporting\_event\_ticket\_info" and click **Create**.

Create view

Views are updated each time you run a query

Name

Cancel

Create

10. Copy the following SQL syntax into the New Query 3 tab and click **Run Query**.

```
SELECT
  sport,
  count(distinct location) as locations,
  count(distinct event_id) as events,
  count(*) as tickets,
  avg(ticket_price) as avg_ticket_price
FROM sporting_event_ticket_info
GROUP BY 1
ORDER BY 1;
```

You query returns two results in approximately five seconds. The query scans 25MB of data, which prior to converting to parquet, would have been 1.59GB of CSV files.

The screenshot shows the AWS Athena Query Editor interface. At the top, there are tabs for 'Athena', 'Query Editor' (which is active), 'Saved Queries', 'History', 'AWS Glue Data Catalog', and 'Workgroup : primary'. Below the tabs, there are three query tabs: 'New query 1', 'New query 2', and 'New query 3' (which is selected). The SQL query is pasted into the editor area. Below the query, there are buttons for 'Run query', 'Save as', and 'Create'. To the right of these buttons, it says '(Run time: 6.21 seconds, Data scanned: 25.97 MB)'. Below the buttons, there is a note: 'Use Ctrl + Enter to run query, Ctrl + Space to autocomplete'. At the bottom, there is a 'Results' section showing a table with 5 columns: 'sport', 'locations', 'events', 'tickets', and 'avg\_ticket\_price'. The table has 2 rows of data.

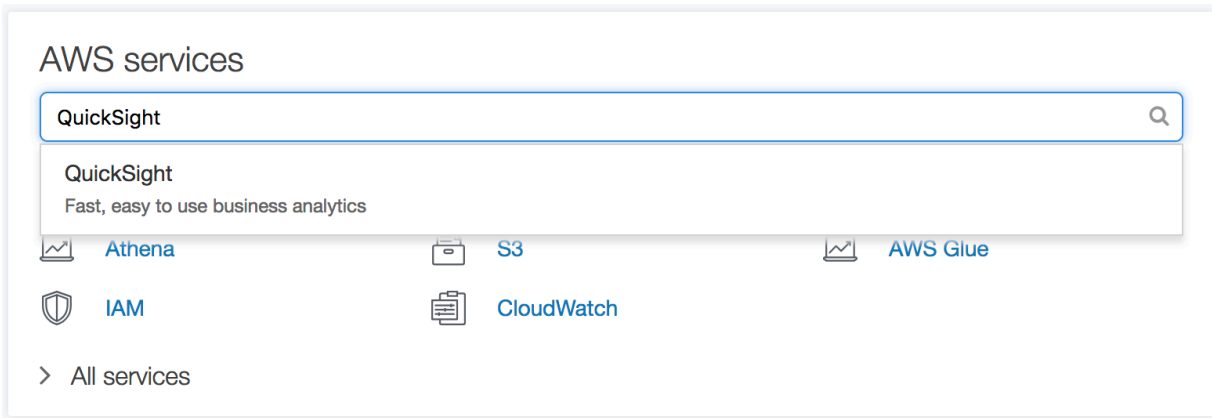
	sport	locations	events	tickets	avg_ticket_price
1	baseball	30	294	958680	53.89345581425812
2	football	25	113	810304	57.40977502271104



## Build an Amazon QuickSight Dashboard ( Proceed to Sagemaker Lab Instead)

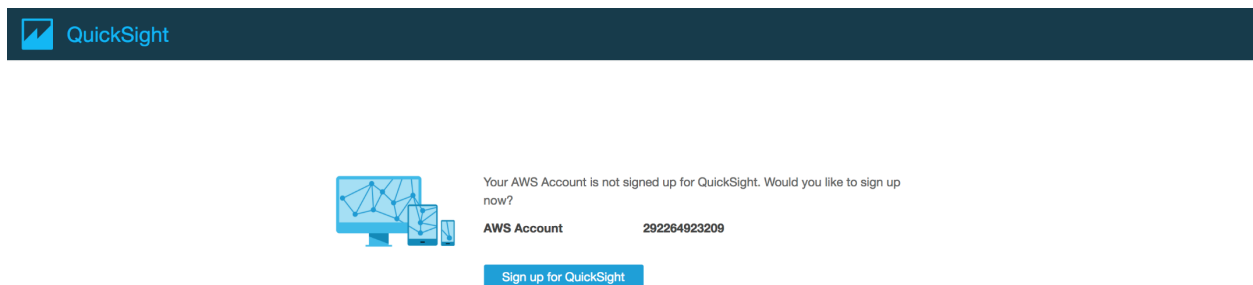
### Set up QuickSight

1. In the AWS services console, search for **QuickSight**.



If this is the first time you have used QuickSight, you are prompted to create an account.

2. Click **Sign up for QuickSight**.



3. For account type, please choose **Enterprise Version**
4. Click **Continue**.
5. On the Create your QuickSight account page, fill out **your name and email address**.
6. Select region and the check boxes to enable **autodiscovery, Amazon Athena, and Amazon S3**.
7. In Select Amazon S3 buckets page, tick your **DMS bucket** (e.g., "xx-dmslabs3buckt").
8. Click **Finish**.

**QuickSight**

## Create your QuickSight account

Edition Standard

**QuickSight account name**  
 ⓘ  
 You will need this for you and others to sign in.

**Notification email address**  
  
 For QuickSight to send important notifications.

**QuickSight capacity region**  
 ⓘ  
 Select a region.

☒ Enable autodiscovery of data and users in your Amazon Redshift, Amazon RDS and AWS IAM services.  
☒ **Amazon Athena**  
Enables QuickSight access to Amazon Athena databases  
Please ensure the right Amazon S3 buckets are also enabled for QuickSight.

☒ **Amazon S3 (1 bucket)**  
Enables QuickSight to auto-discover your Amazon S3 buckets [Choose S3 buckets](#)  
☐ **Amazon S3 Storage Analytics**  
Enables QuickSight to visualize your S3 Storage Analytics data  
☐ **Amazon IoT Analytics**  
Enable QuickSight to visualize your IoT Analytics data

[Finish](#)

9. On the QuickSight landing page, click **Manage Data**.

10. Click **New Data Set** button.

11. On the Create a Data Set page, select **Athena** as the data source.

**QuickSight**

[Data sets](#) 14.5MB of SPICE used of 1GB in N. Virginia

### Create a Data Set

FROM NEW DATA SOURCES

**Upload a file**  
(.csv, .tsv, .cll, .elf, .xlsx, .json)

**Salesforce**  
Connect to Salesforce

**S3 Analytics**

**S3**

**Athena**

**RDS**

12. For Data source name, type "ticketdata-qs" and click **Validate connection**.

13. Click **Create data source**.

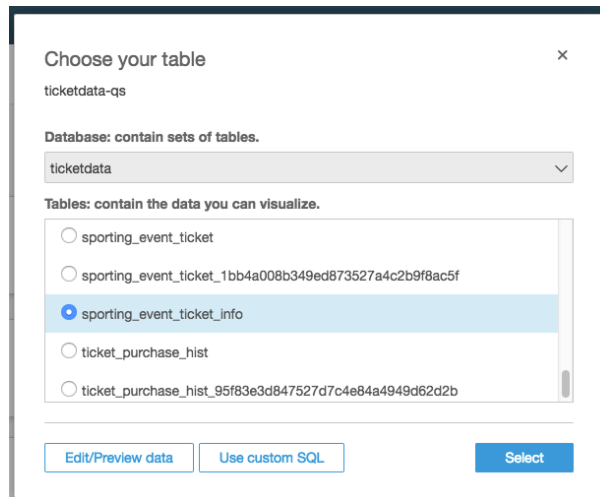
**New Athena data source** ⓘ

**Data source name**

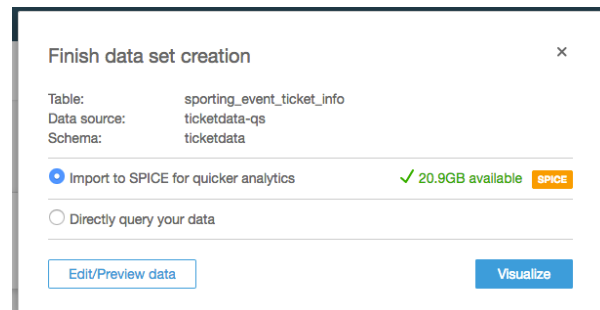
✓ Validated
SSL is enabled

[Create data source](#)

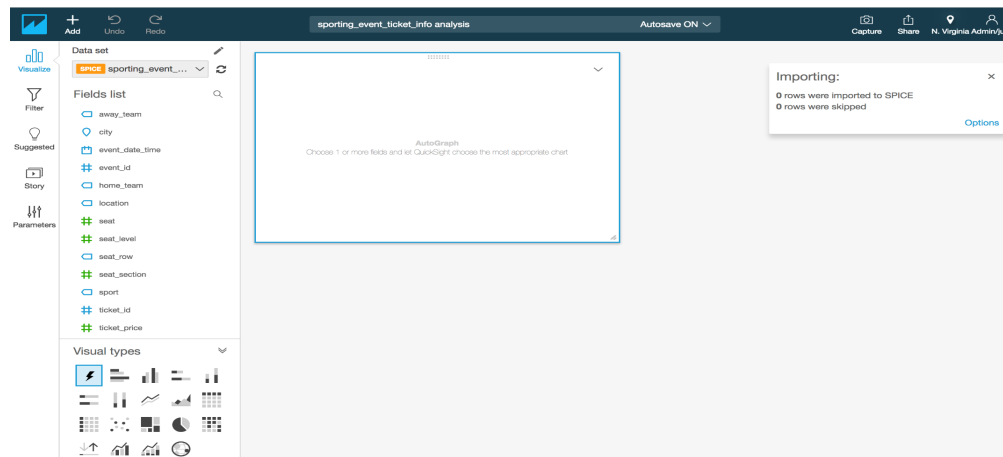
14. In the Database drop-down list, select the database name you created in the AWS Glue lab.
15. Choose the "sporting\_event\_ticket\_info" table and click **Select**.



16. To finish data set creation, choose the option **Import to SPICE** for quicker analytics and click **Visualize**.



You will now be taken to the QuickSight Visualize interface where you can start building your dashboard.

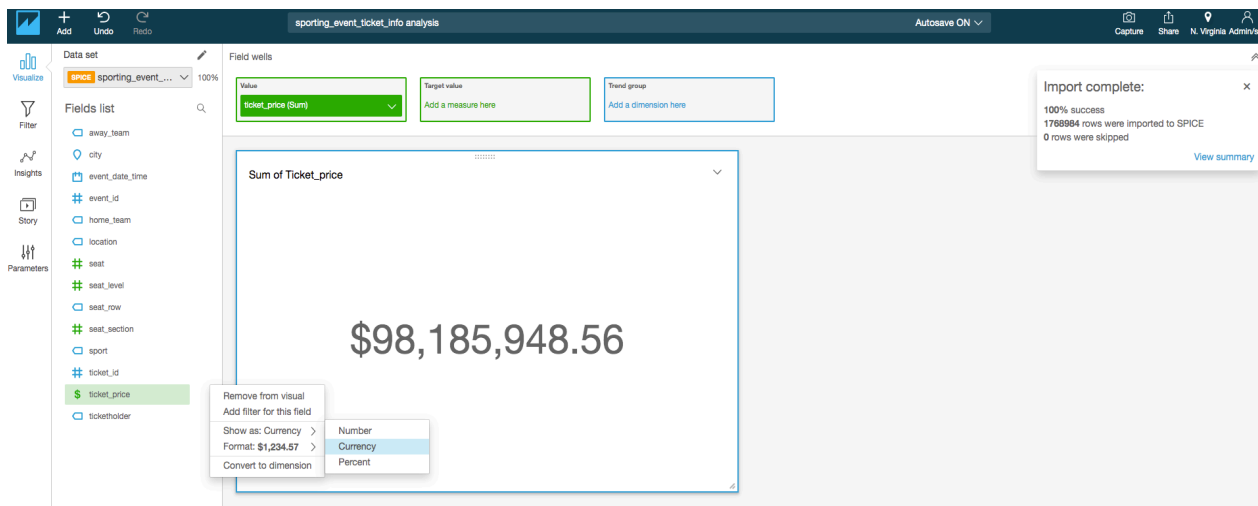


**Note:** The SPICE dataset will take a few minutes to be built, but you can continue to create some charts on the underlying data.

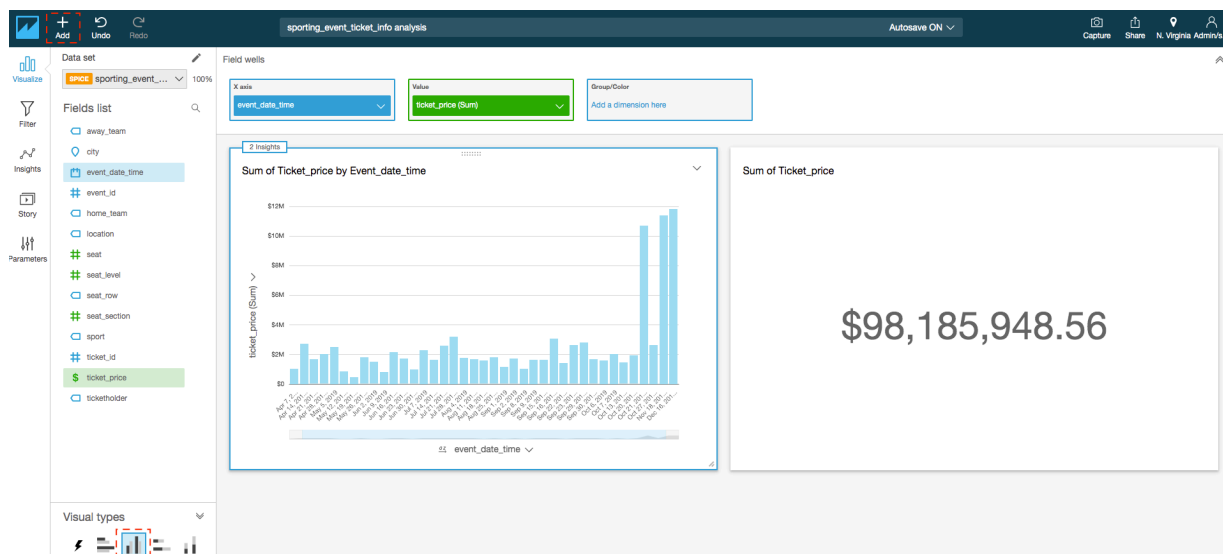
## Create QuickSight Charts

In this section we will take you through some of the different chart types.

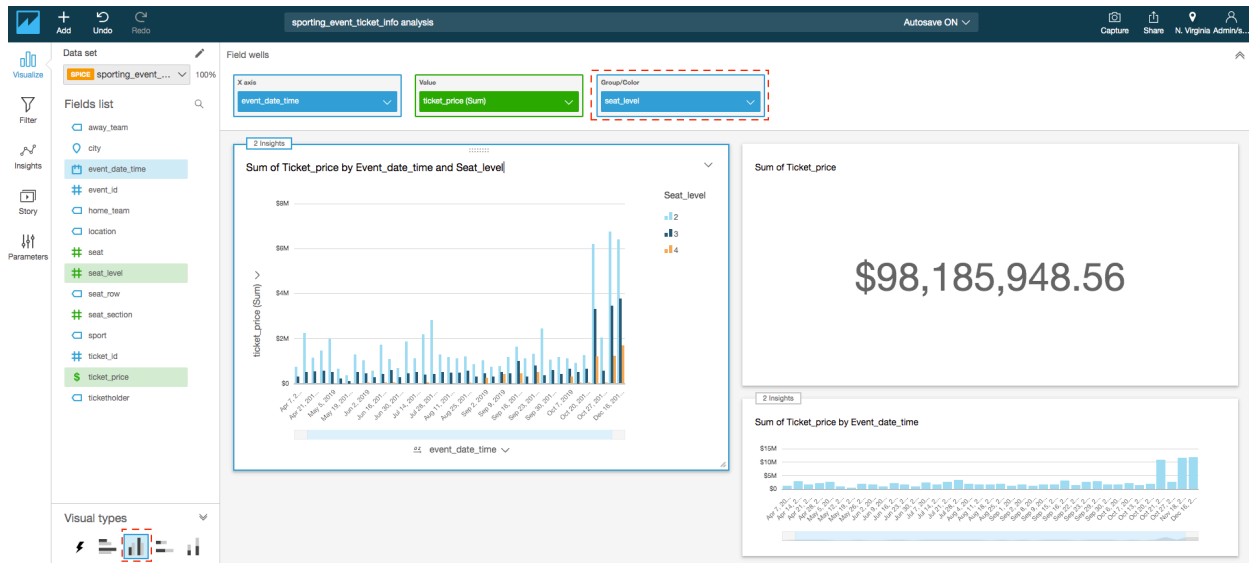
1. In the Fields list, click the "ticket\_price" column to populate the chart.
2. Click the expand icon in corner of "ticket\_price" field and select format as currency to show numbers in dollar amount.



3. You can add a new visual and keep building your dashboard by clicking **Add** button at top left corner of screen. In the **Visual types** area, choose the **Vertical bar chart** icon. This layout requires a value for the X-axis. Click the "event\_date\_time" field and you should see the visualization update.

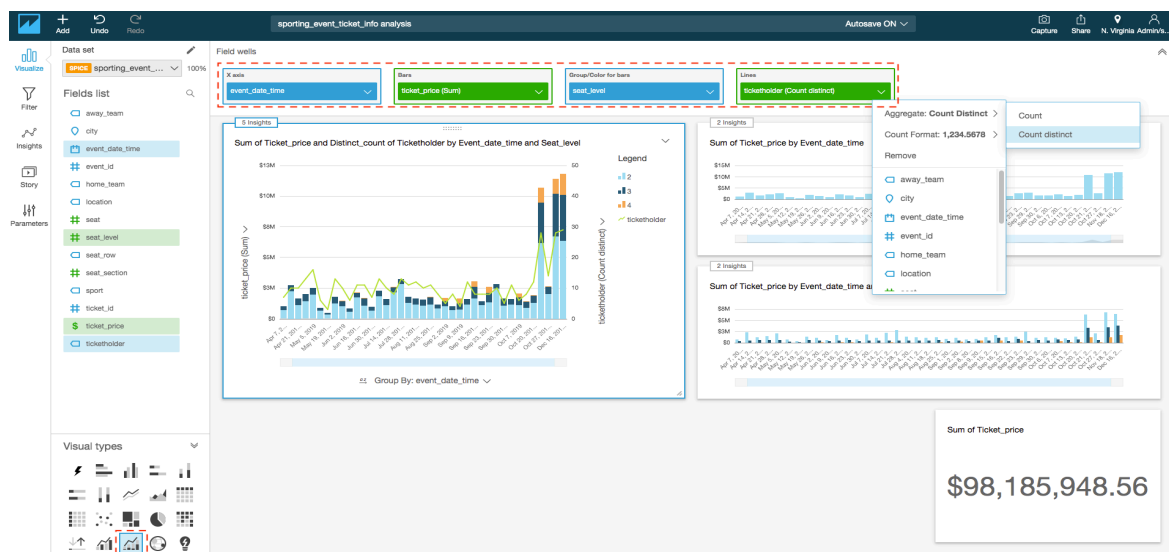


4. Add a third new Visual and you can drag and move other visuals to adjust space in dashboard. In the Fields list, click and drag the **seat\_level** field to the **Group/Color** box in the Field wells pane. You can also use the slider below the x axis to fit all of the data.

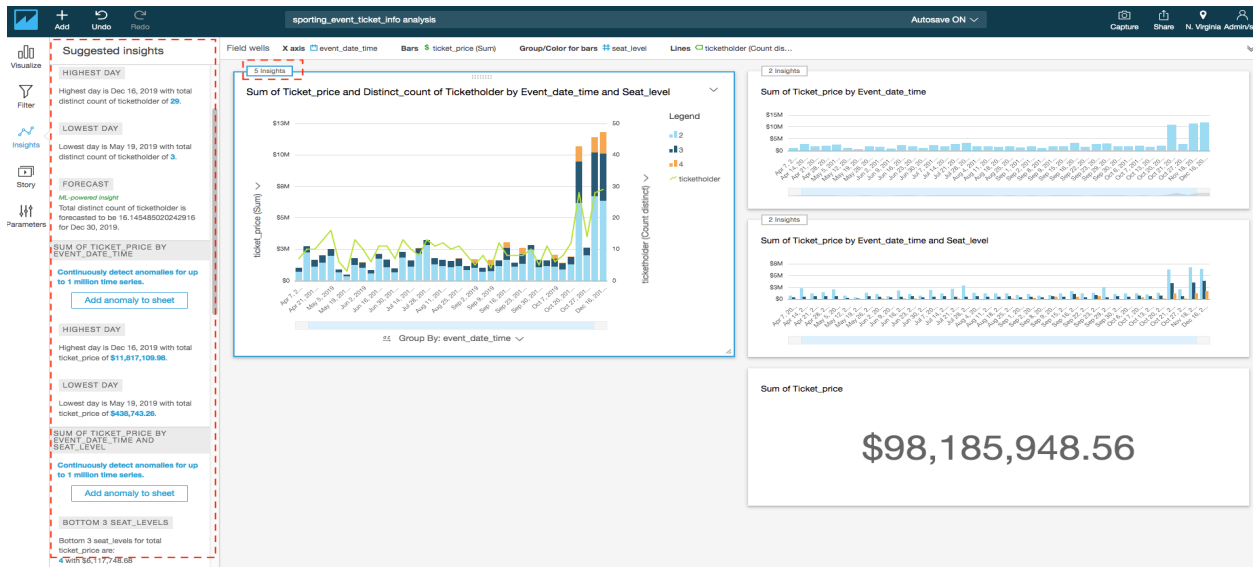


Let's build on this one step further by changing the chart type to "**Clustered bar combo chart**" and adding in the ticketholder for the Lines.

5. In the Visual types area, choose the **Clustered bar combo chart** icon.
6. In the Fields list, click and drag the **ticketholder** field to the **Lines** box in the Field wells pane.



7. In the Field wells pane, click the Lines box and choose **Count Distinct** for Aggregate. You can then see the **y-axis** update on the **right-hand** side.
8. Click on **insight** icon on top of each chart and explore insight information in right hand pan in simple English.

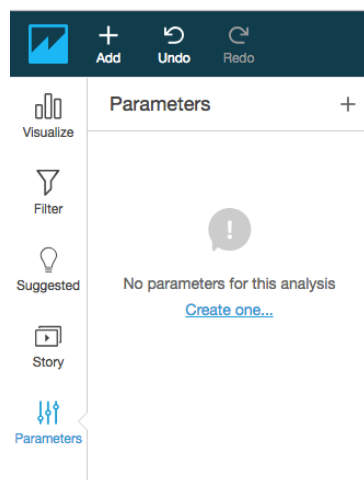


Feel free to experiment with other chart types and different fields to get a sense of the data.

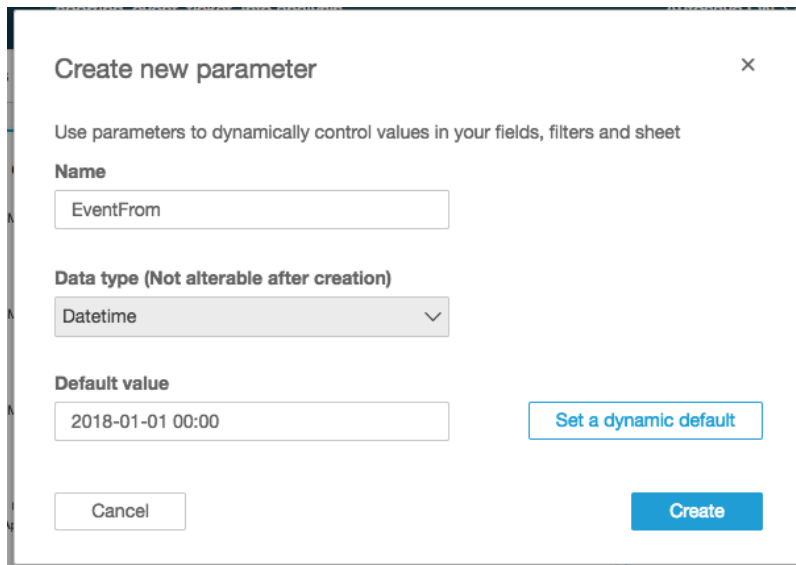
## Create QuickSight Parameters

In the next section we are going to create some parameters with controls for the dashboard, then assign these to a filter for all the visuals.

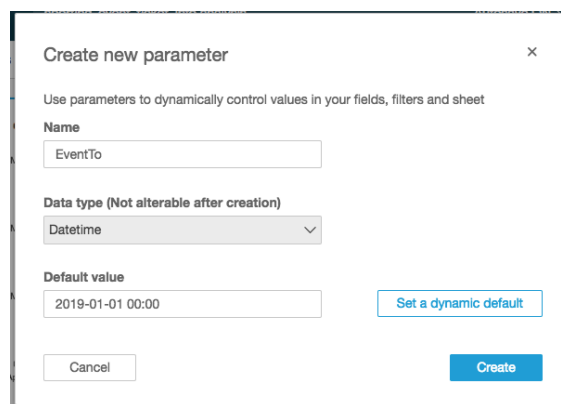
1. In the left navigation menu, select **Parameters**.



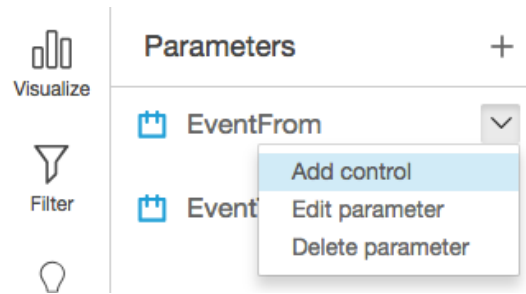
2. Click **Create one** to create a new parameter with a Name.
3. For Name, type **EventFrom**.
4. For Data type, choose **Datetime**.
5. For Default value, select the value from calendar as start date available in your graph for event\_date\_time. For example, **2018-01-01 00:00**.
6. Click **Create**, and then close the Parameter Added dialog box.



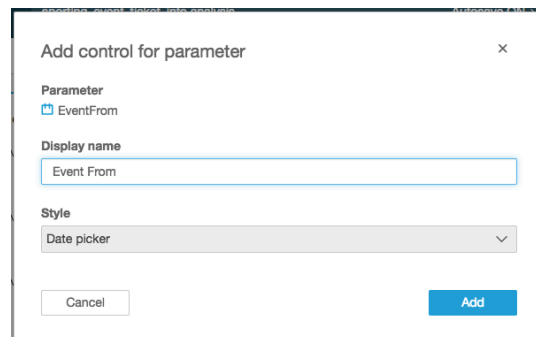
7. Create another parameter with the following attributes:
  - a. Name: EventTo
  - b. Data type: Datetime
  - c. For Default value, select the value from calendar as end date available in your graph for event\_date\_time. For example, 2019-01-01 00:00



8. Click **Create**.
9. In the Parameter Added dialog box, click **Filter** and then click **Close**.
10. Click the drop-down menu for the EventFrom parameter and choose **Add control**.

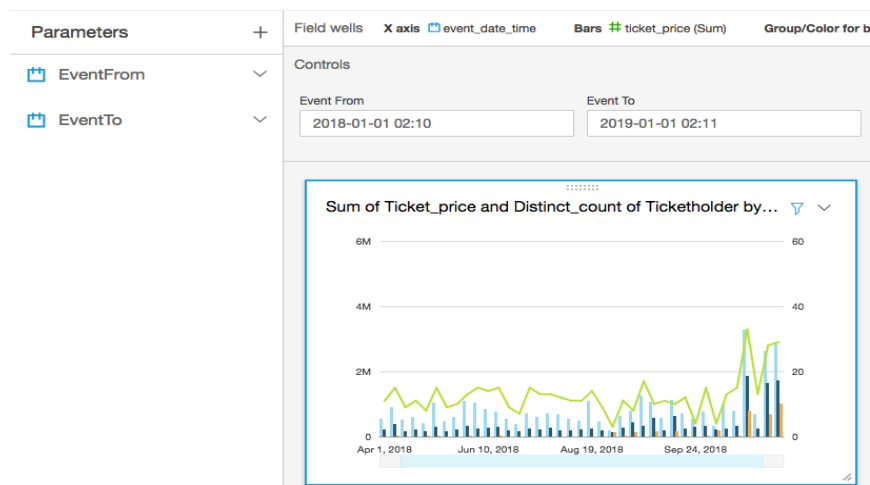


11. For Display name, specify **Event From** and click **Add**.



12. Repeat the process to add a control for **EventTo** with display name **Event To**.

You should now be able to see and expand the Controls section above the chart.

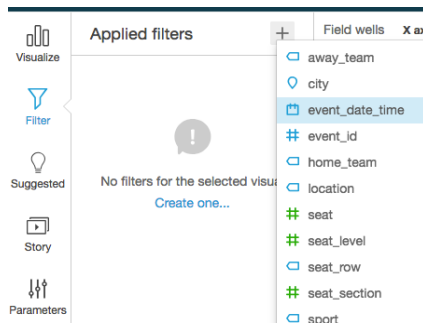


## Create a QuickSight Filter

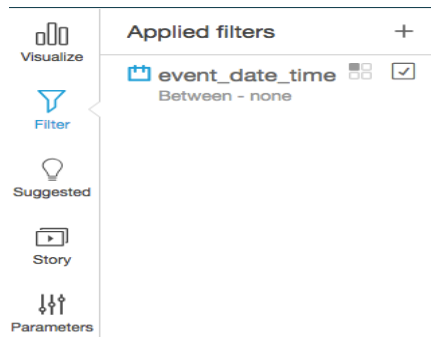
To complete the process, we will wire up a filter to these controls for all visuals.

1. In the left navigation menu, choose **Filter**.
2. Click the plus icon (+) to add a filter for the field "event\_date\_time".

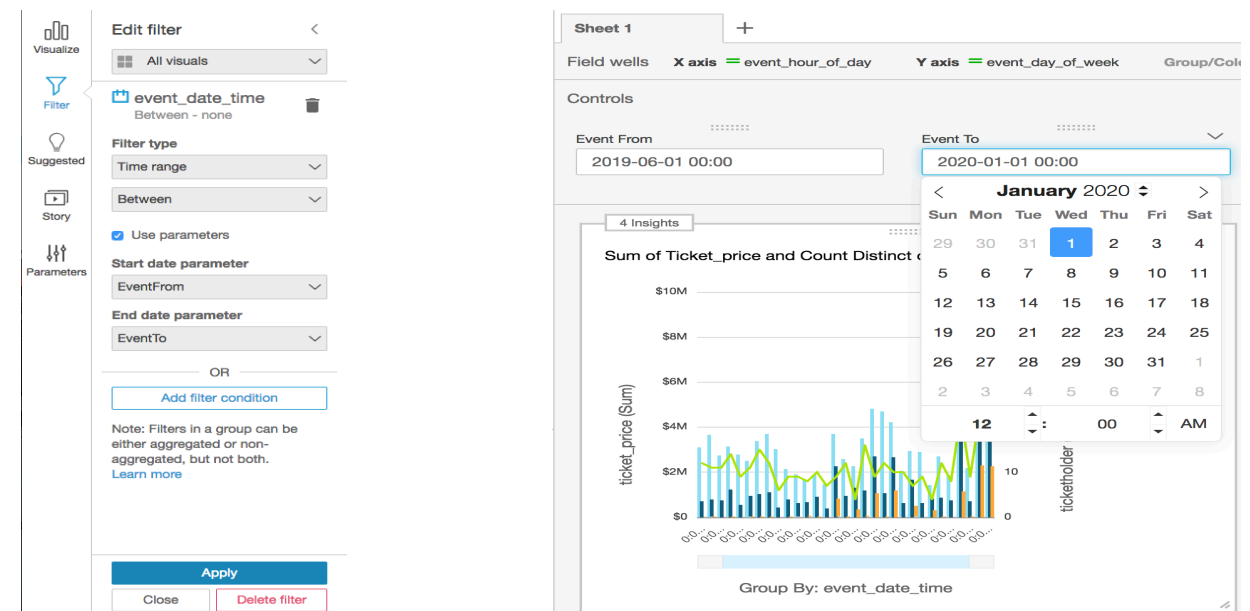




3. Click this filter to edit the properties.



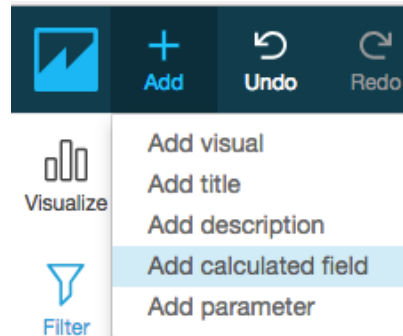
4. Choose to make this filter apply to **All visuals**.
5. For Filter type, choose **Time range** and **Between**.
6. Select option Use **Parameter**.
7. For Start date parameter, choose **EventFrom**.
8. For End date parameter, choose **EventTo**.
9. Click **Apply**.
10. Change the report filter Event From: 2019-06-01, Event To: 2020-01-01



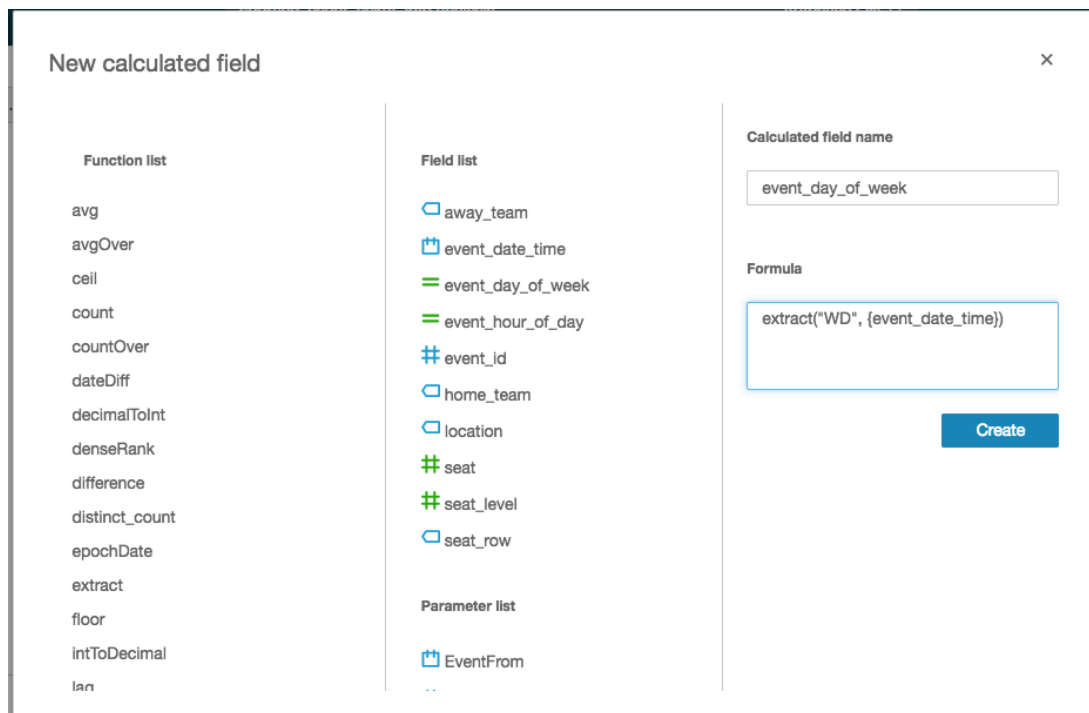
## Add Calculated Fields

In the next section, you will learn how to add calculated fields for "day of week" and "hour of day" to your dataset and a new scatter plot for these two dependent variables.

1. Click the Add button on the top left and select **Add a calculated field**.



2. For **Calculated field name** type "event\_day\_of\_week".
3. For Formula, type `extract("WD", {event_date_time})`  
Note: extract returns a specified portion of a date value. Requesting a time-related portion of a date that doesn't contain time information returns 0. WD: This returns the day of the week as an integer, with Sunday as 1.
4. Click **Create**.



New calculated field

Function list

- avg
- avgOver
- ceil
- count
- countOver
- dateDiff
- decimalToInt
- denseRank
- difference
- distinct\_count
- epochDate
- extract
- floor
- intToDecimal
- lan

Field list

- away\_team
- event\_date\_time
- event\_day\_of\_week
- event\_hour\_of\_day
- event\_id
- home\_team
- location
- seat
- seat\_level
- seat\_row

Parameter list

- EventFrom

Calculated field name

event\_day\_of\_week

Formula

extract("WD", {event\_date\_time})

Create

5. Add another calculated field with the following attributes:
  - a. Calculated field name: "event\_hour\_of\_day"
  - b. Formula: `extract("HH", {event_date_time})`

Note: HH: This returns the hour portion of the date.

**New calculated field**

**Function list**

- addDateTime
- cell
- coalesce
- concat
- dateDiff
- decimalToInt
- epochDate
- extract**
- floor
- formatDate
- ifelse
- intToDecimal
- isNotNull
- isNull
- left

**Field list**

- away\_team
- event\_date\_time
- event\_id
- event\_week\_day
- home\_team
- location
- seat
- seat\_level
- seat\_row
- seat\_section
- sport
- ticket\_id
- ticket\_price
- ticketholder

**Calculated field name**

event\_hour\_of\_day

**Formula**

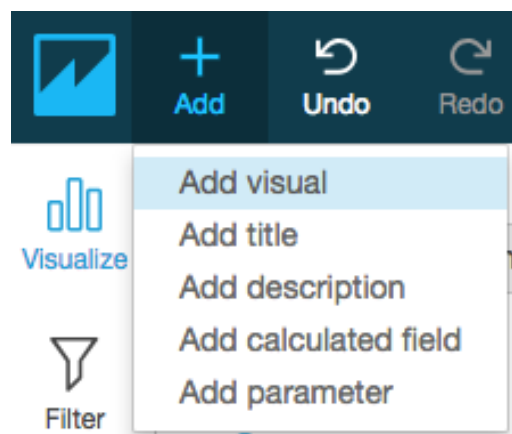
`extract("HH", {event_date_time})`

**extract** extracts a specified datepart from a date value. datepart can be one of "YYYY", "MM", "DD", "HH", "MI", "SS", "WD"

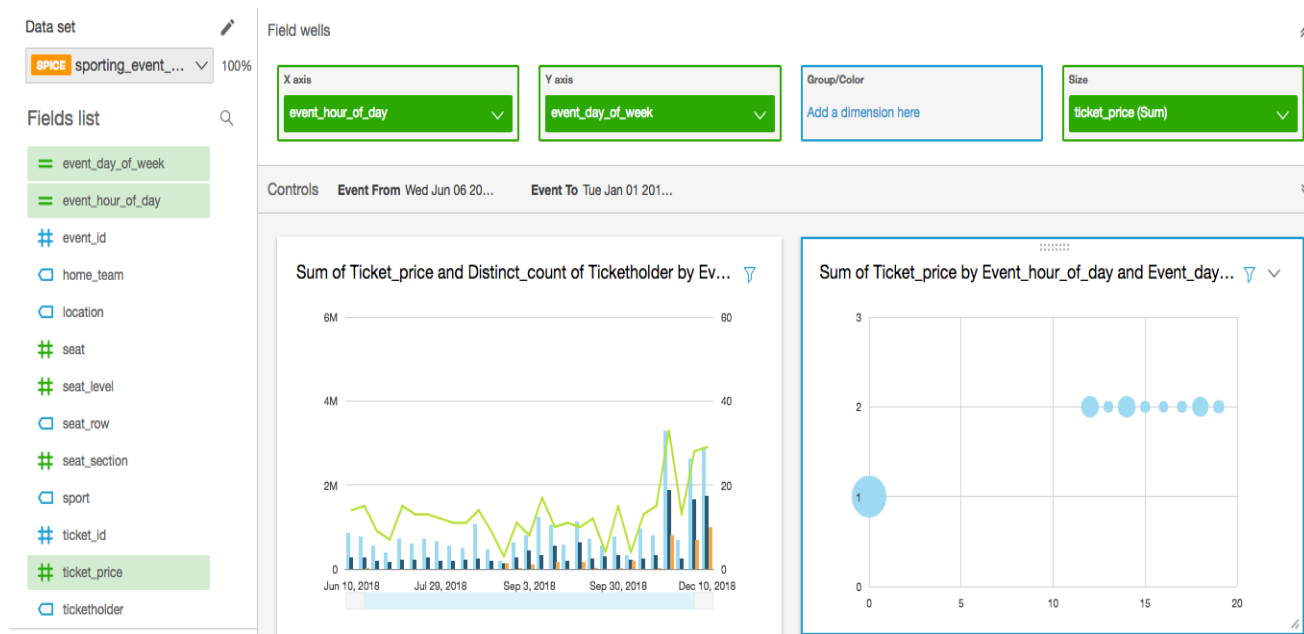
**Syntax:** `extract(datepart, date)`

**Create**

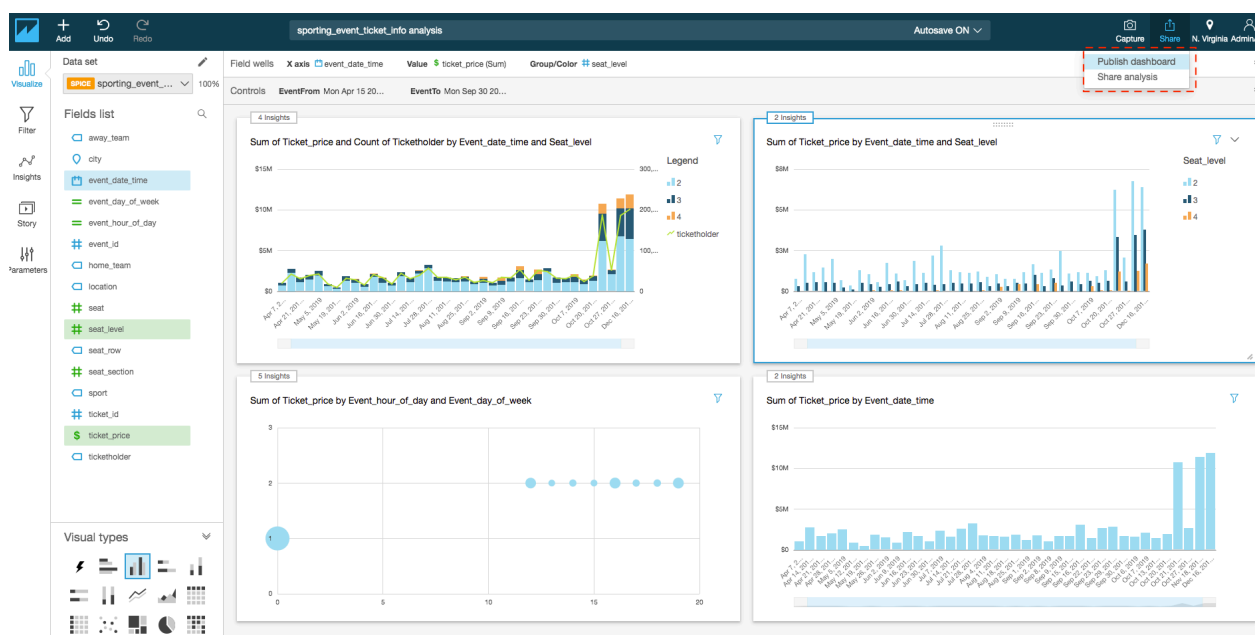
6. Click Add button in the top left and choose **Add visual**.



7. For field type, select the scatter plot.
8. In the Fields list, select and drag the following attributes to the Field wells pane to set the graph attributes:
  - a. X-axis: "event\_hour\_of\_day"
  - b. Y-axis: "event\_day\_of\_week"
  - c. Size: "ticket\_price"



Since now you have completed your dashboard then you can publish it by clicking on top right corner of screen.



A *dashboard* is a read-only snapshot of an analysis that you can share with other Amazon QuickSight users for reporting purposes. In Dashboard other users can still play with visuals and data but that will not modify dataset.

You can share an analysis with one or more other users with whom you want to collaborate on creating visuals. Analysis provide other uses to write and modify data set.

## Amazon QuickSight ML-Insights (Optional)

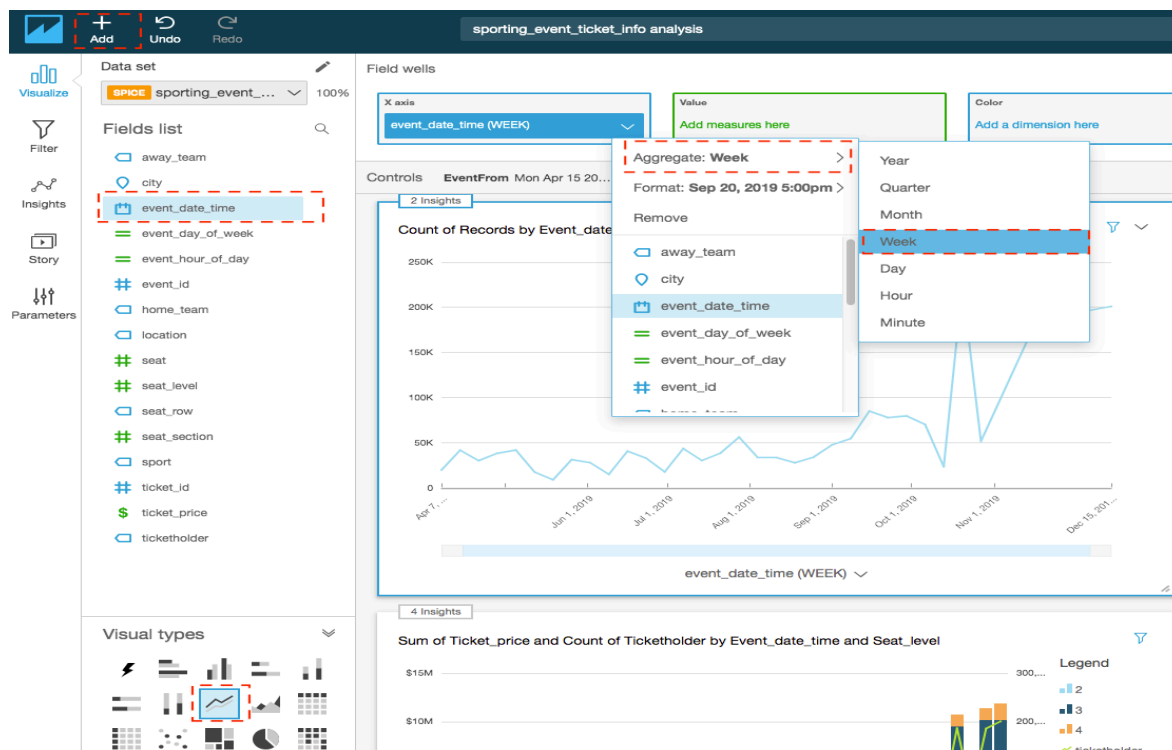
With Amazon QuickSight, you can add Machine Learning capabilities to your visuals, easily, with one click action. There are 3 types of Machine Learning Insights

- Narrative
- Anomaly Detection
- Forecasting

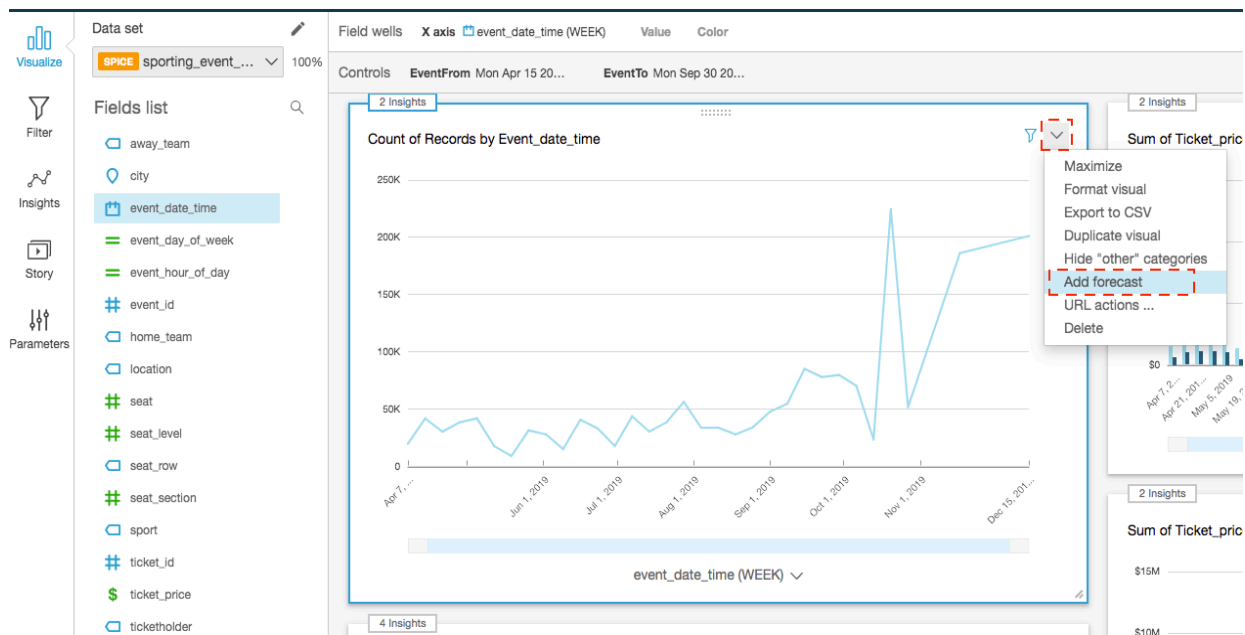
ML-Insights is only available to enterprise version of QuickSight. You will need to upgrade to Enterprise Edition before you start with the task. To upgrade your Amazon QuickSight Subscription from Standard Edition to Enterprise Edition please follow this guide <https://docs.aws.amazon.com/quicksight/latest/user/upgrading-subscription.html>

Let's see how we can add a bit of forecasting in our dashboard. Forecasting works with timeseries, which is better represented with a line graph. Let's first create a line graph.

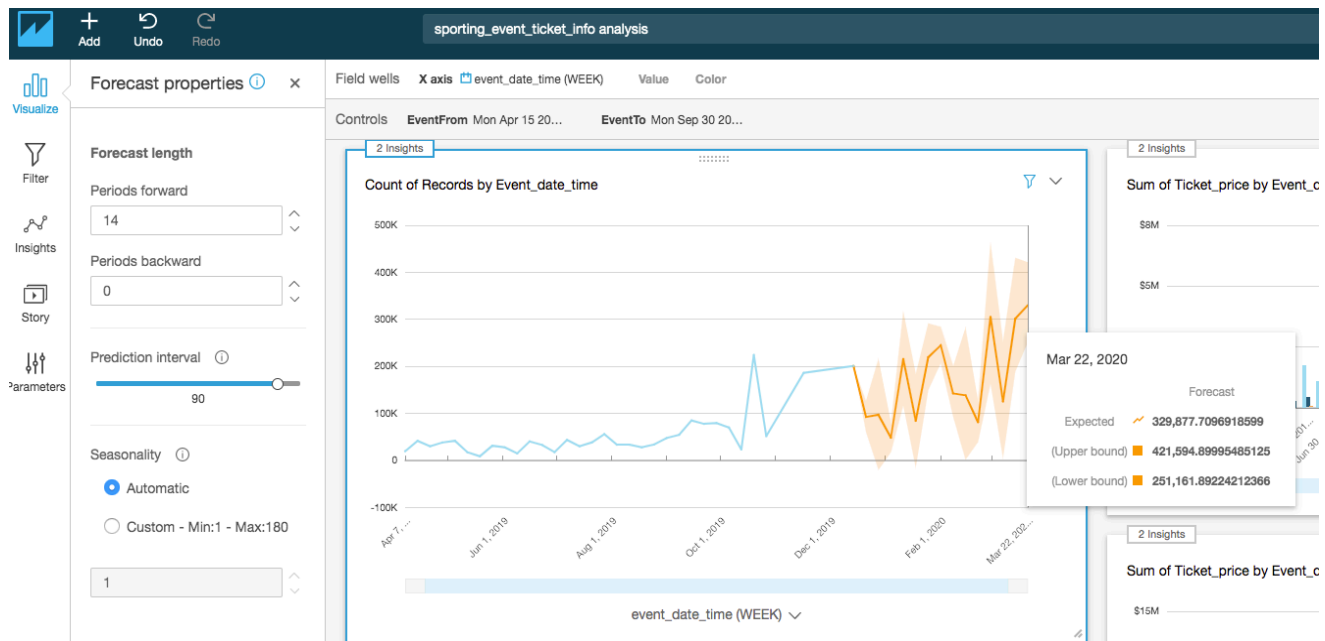
1. Click **add Visual** at top left corner of screen, and select **Line Chart** and add the **event\_date\_time** as the x-axis and **aggregate by week**. As shown in below screenshot



2. Add forecasting to the visual. To do that, click on the drop-down arrow on the top right corner of the visual, and then click **Add forecast**.



The visual will add forecast, you can hover over and explore forecasted data as shown below:



Feel free to explore with the properties of the forecast algorithm.