Visualizing Big Data at Scale with Databricks and Looker

Table of Contents

Overview

Recent years of Big Data adoption has enabled organizations to analyze more complex models and deliver more accurate insights. Spark has become the de-facto platform for high scale analytics, solving complex analytical problems in industries such as finance and healthcare. The Databricks Unified Analytics Platform, created by the inventors of Spark, uses Spark at the core, but includes performance, security and collaboration enhancements that make it the easiest, fastest way to get the massive insights Spark is capable of. In the greater enterprise, these insights need to be made available to a larger audience, less technical and more expert in business in nature, and Looker has proven a leader for enabling this.

Since Looker enables data exploration and transformation by querying data where it lives - analytics directly benefits from Databricks' powerful compute engine and querying power. For all the reasons Databricks has become a natural choice for large scale data processing, Looker can provide a complementary exploration and visualization layer because provides direct access to the analytics models and results.

Databricks enables access to it's power with the easy Spark SQL interface. Users can take advantage of Databricks' performance to analyze structured data in a familiar and efficient way using a SQL-like interface. It can also demonstrate the clear advantages of Schema-on-Write versus Schema-on-Read: the ability to copy and query data in its native format, thereby enabling ETL on the fly, has obvious advantages for analyzing data at large scale.

SQL has effectively proven itself as the natural language for data analysis - favored by analysts to easily write productive queries and is now supported across the big data ecosystem. The Looker platform being able to write SQL to a data source already optimized for performance, like Spark SQL, enables analysis to happen at the most granular level, without having to utilize a complex ETL process where pre-aggregated data is moved elsewhere for exploration and visualization. The underlying power of Databricks can be used in an intuitive and powerful way to transform data at the time of query, such that row level of data can be examined.

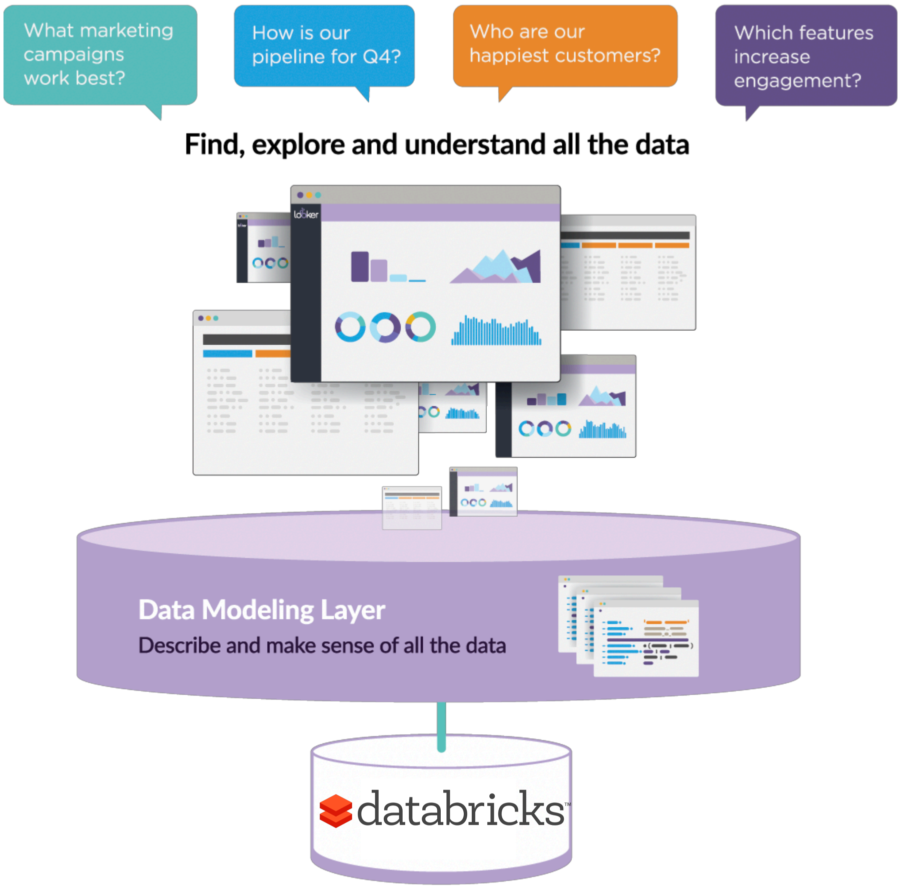
Together, Databricks and Looker allow your organization to query and analyze large datasets without sacrificing self-service and agility for data scientists and business analysts.

How it works

Databricks is a Spark-based cloud platform for data engineering and data science, that enables organizations to create a large scale analytics infrastructure in seconds.

Looker is a lightweight application that can be installed on-premise or in the cloud. Once configured, Looker can connect to Databricks via Spark SQL using a standard JDBC connection. The data is accessed where it lives, alleviating the need to summarize or warehouse and providing analytics as real-time and the data source.

How Looker works with Databricks


What marketing campaigns work best?
How is our pipeline for Q4?
Who are our happiest customers?
Which features increase engagement?
Find, explore and understand all the data
looker
Data Modeling Layer
Describe and make sense of all the data
databricks

Getting Started

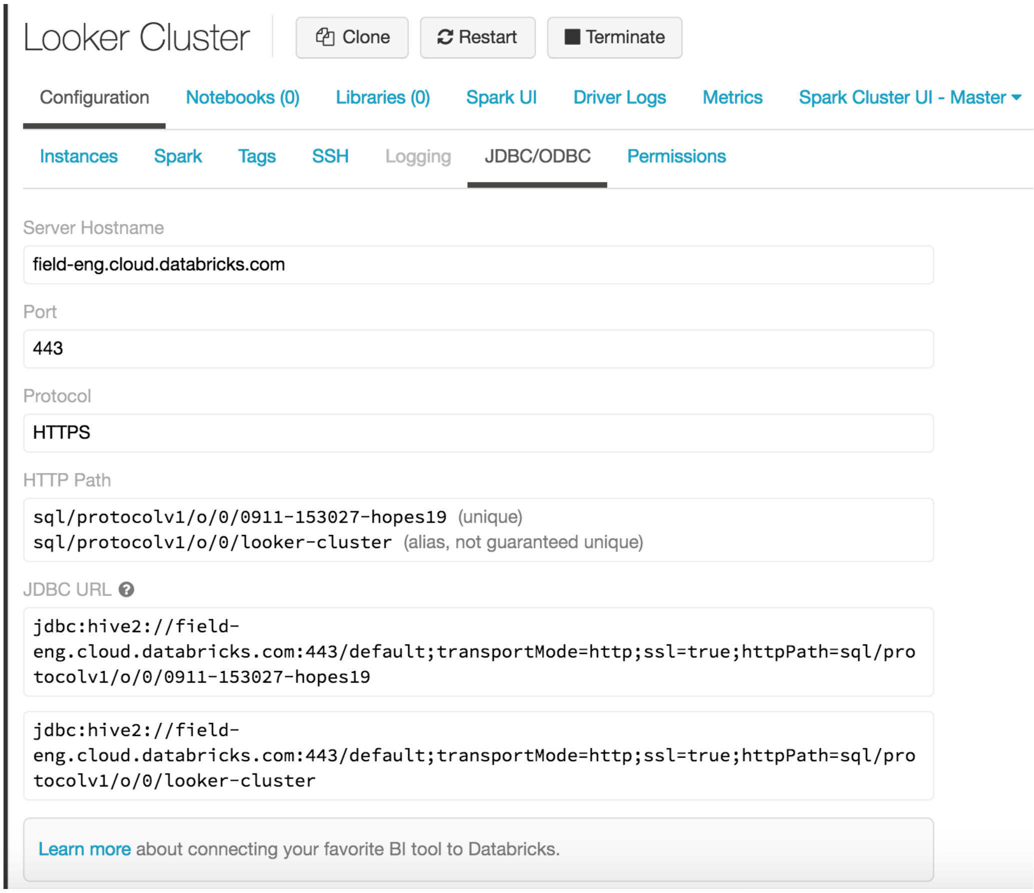
You can begin using Databricks by going to https://databricks.com/try-databricks
You can begin using Looker by going to https://looker.com/free-trial

Implementation Guide

Connecting Looker to a Databricks Cloud Cluster

Step 1: Get Cluster Information

1. Create a Spark Cluster in Databricks. Documentation here.
2. Once the cluster is up and running select the 'JDBC/ODBC' tab under cluster configuration. This information will be used to create a connection in Looker



Server Hostname

field-eng.cloud.databricks.com

Port

443

Protocol

HTTPS

HTTP Path

```
sql/protocolv1/o/0/0911-153027-hopes19  (unique)
sql/protocolv1/o/0/looker-cluster  (alias, not guaranteed unique)
```

JDBC URL

```
jdbc:hive2://field-eng.cloud.databricks.com:443/default;transportMode=http;ssl=true;httpPath=sql/protocolv1/o/0/0911-153027-hopes19
```

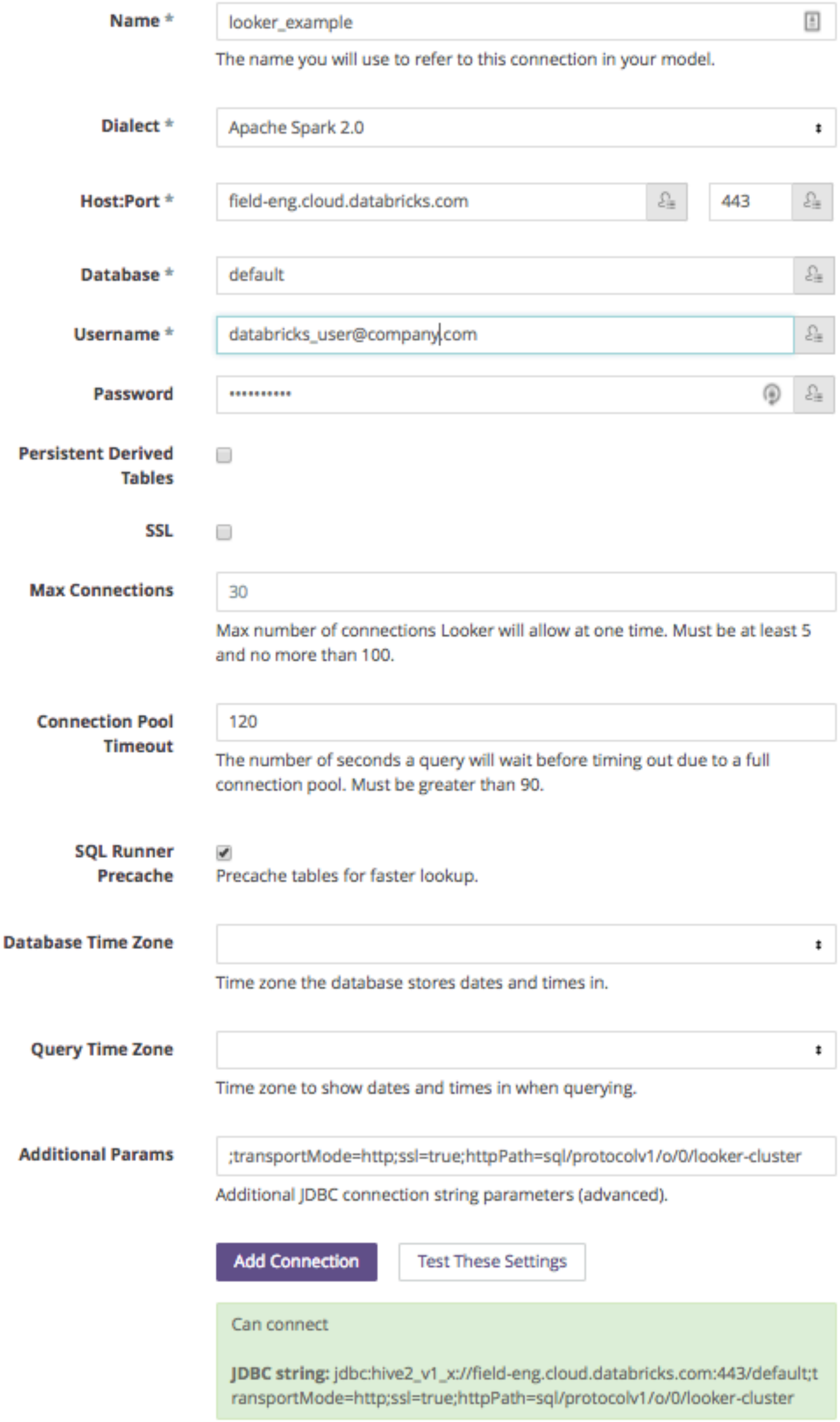
```
jdbc:hive2://field-eng.cloud.databricks.com:443/default;transportMode=http;ssl=true;httpPath=sql/protocolv1/o/0/looker-cluster
```

Learn more about connecting your favorite BI tool to Databricks.

Step 2: Connect Looker to your Databricks Cloud Cluster

1. Go to Admin -> Connections -> New Connection

2. Fill in the connection parameters:
    - Enter your connection name
    - Select ‘Apache Spark 2.0’ as the Dialect
    - Enter the Server Hostname found in JDBC/ODBC config as the Host
    - Set the port to the value defined in JDBC/ODBC config (should be 443)
    - Select ‘default’ as the database
    - Enter Databricks Username and Password
    - Don’t enable PDTs
    - Don’t enable SSL. SSL is required to connect to Databricks, but will be included by default in Additional Params.
    - Leave Max Connections and Connection Pool Timeout at their defaults
    - Leave SQL Runner Precache checked
    - Leave Database Time Zone blank (assuming you are storing everything in UTC)
    - Adjust Query Time Zone if you want to translate queries into other time zones
    - In Additional Parameters copy the second half of the JDBC URL defined in the JDBC/ODBC config starting at ‘;transportMode’ (must include preceding ‘;’)
    - Click ‘Test These Settings’ to make sure that you have everything set properly
3. Click Add Connection

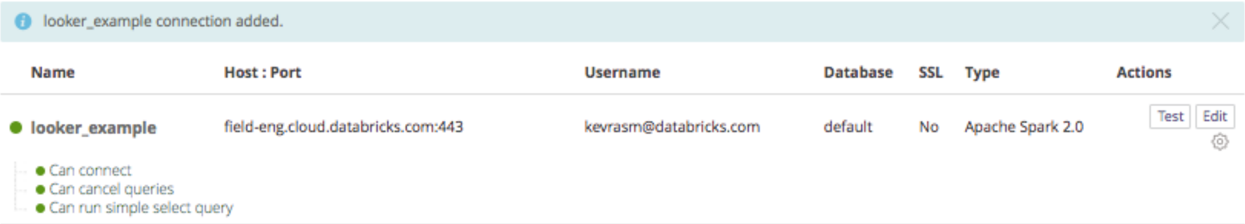


Name *	looker_example The name you will use to refer to this connection in your model.
Dialect *	Apache Spark 2.0
Host:Port *	field-eng.cloud.databricks.com : 443
Database *	default
Username *	databricks_user@company.com
Password	•••••••••••
Persistent Derived Tables	□
SSL	□
Max Connections	30 Max number of connections Looker will allow at one time. Must be at least 5 and no more than 100.
Connection Pool Timeout	120 The number of seconds a query will wait before timing out due to a full connection pool. Must be greater than 90.
SQL Runner Precache	☑ Precache tables for faster lookup.
Database Time Zone	 Time zone the database stores dates and times in.
Query Time Zone	 Time zone to show dates and times in when querying.
Additional Params	;transportMode=http;ssl=true;httpPath=sql/protocolv1/o/0/looker-cluster Additional JDBC connection string parameters (advanced).

Can connect

JDBC string: jdbc:hive2_v1_x://field-eng.cloud.databricks.com:443/default;transportMode=http;ssl=true;httpPath=sql/protocolv1/o/0/looker-cluster

Once you have added the connection, it will display along side connections to other data sources.



looker_example connection added.

Name	Host : Port	Username	Database	SSL	Type	Actions
● **looker_example**	field-eng.cloud.databricks.com:443	kevrasm@databricks.com	default	No	Apache Spark 2.0	Test Edit

- ● Can connect
- ● Can cancel queries
- ● Can run simple select query

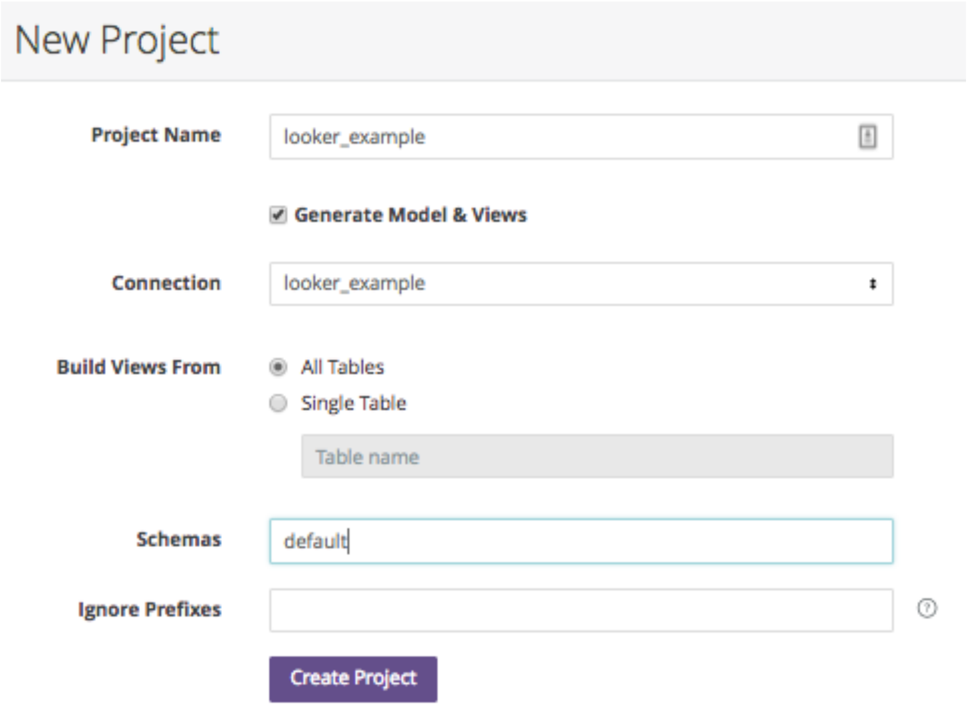
Step 3: Begin modeling your database by creating a project and running the generator

Note that this step assumes that there are permanent tables stored in the default database of your cluster

1. If necessary get into “Developer Mode” by clicking the `Dev` button from **OFF** to **ON**
2. Go to LookML -> Manage Projects
3. Click on New LookML Project
4. Configure the new project

- Give the project a name
- Select the Connection name that you used in Step 3
- Select All Tables
- Use `default` as the Schemas, unless you have other databases to model in the cluster

1. Click Create Project



New Project

Project Name: looker_example

☑ Generate Model & Views

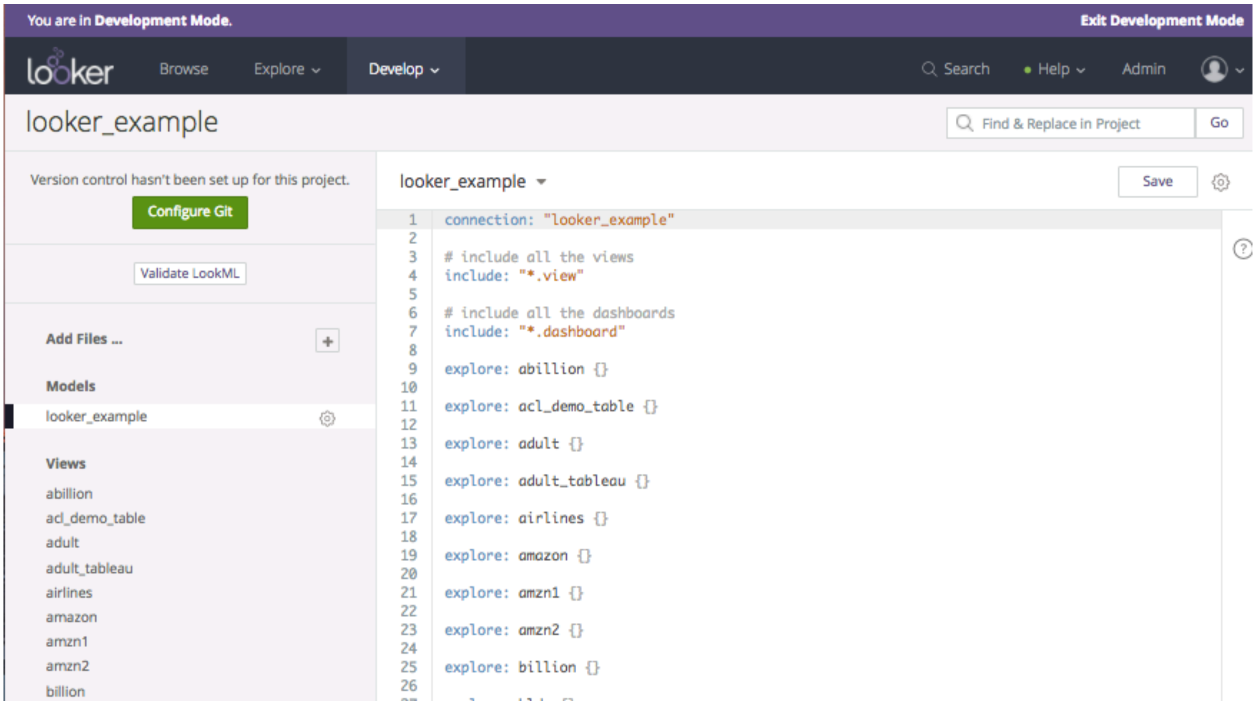
Connection: looker_example

Build Views From: ● All Tables ○ Single Table

Schemas: default

Ignore Prefixes:

After creating the project and the generator runs, you will see something like the following:



Version control hasn't been set up for this project.

```
connection: "looker_example"

# include all the views
include: "*.view"

# include all the dashboards
include: "*.dashboard"

explore: abillion {}

explore: acl_demo_table {}

explore: adult {}

explore: adult_tableau {}

explore: airlines {}

explore: amazon {}

explore: amzn1 {}

explore: amzn2 {}

explore: billion {}
```

You will then find one model file and multiple view LookML files. The model file shows the tables in the schema and any discovered join relations between them, and the view files list each dimension (column) available for each table in the schema.

Summary

We have seen customers using these products together to provide an easy and intuitive way for business users to visualize and discover the powerful analytics results of Spark. Using Looker and Databricks, you can experience the following benefits:

- Easy to Use – Make analysts productive instantly through easy to use visualizations
- Fastest User Adoption – Enable widespread use of analytics throughout your organization through fast user adoption
- Process More Data Faster - Provides the fastest implementation of Spark by using Databricks
- Answer Your Toughest Questions - Run the most complex analytics problems, providing answers to your toughest questions. (See this benchmarking study)
- Bigger Insights, More Intuitively - Easy to use on your toughest problems makes bigger insights more intuitive
- Game Changing Insights – Make complex analytical answers available to more users to drive cross-company insights that change the game

Spark is the fastest analytics processing engine available. Databricks provides the Unified Analytics Platform, built on Spark by the inventors of Spark. It is performance tuned to run 7-10 times faster than vanilla spark, plus it has advanced security and content connectors. The Unified Analytics Platform provides a collaborative platform for data scientists, data engineers, and business users. It enables data scientists to create and run analytical models, including machine learning and artificial intelligence. It enables data engineers to run scheduled projects to extract, join and cleanup data. Its performance and flexibility makes ETL one of Databricks' most popular use cases.

When you combine the analytics power of Databricks with the intuitive, user friendly interface of Looker, you provide a way for business users throughout your organization to use analytics to realize deeper insights and make better business decisions.