



**BigDataWorkGroup**



1

Big Data Structures

# DataScience in The Cloud



# An Introduction to Data Science

- ▶ Data Science refers to an emerging area of work concerned with the **collection**, **preparation**, **analysis**, **visualization**, **management** and **preservation** of large collections of information.
  - ▶ An Introduction to Data Science
  - ▶ Jeffrey Stanton
  - ▶ Syracuse University School of Information Studies
- ▶ A data scientist is someone who can **obtain**, **scrub**, **explore**, **model** and **interpret** data, blending **hacking**, **statistics** and **machine learning**. Data Scientists not only are adept at working with data, but appreciate data itself as a first-class product
  - ▶ Hilary Mason, chief scientist at bit.ly
- ▶ Data wrangling, Data jujitsu, Data munging

# Data Products

- ▶ Data science is about building data products, not just answering questions.
- ▶ Data-driven apps : Spellcheckers ,Machine Translator
- ▶ Interactive visualization : Google flu application, Global Burden of Disease
- ▶ Online Databases : Enterprise data warehouse, Sloan Digital Sky Survey

# eScience = Data Science

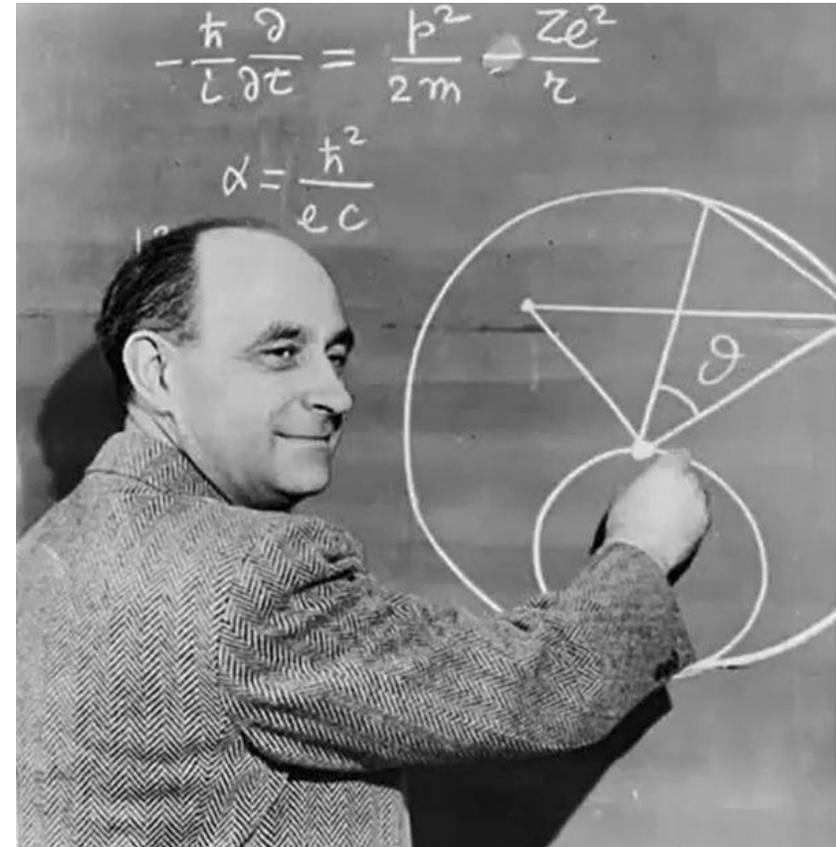
- ▶ Empirical:
  - ▶ Observe Natural world, Replicate natural world in laboratory



*public domain*

# eScience = Data Science

- ▶ Empirical:
  - ▶ Observe Natural world, Replicate natural world in laboratory
- ▶ Theoretical:
  - ▶ to model the empirical observation use theory



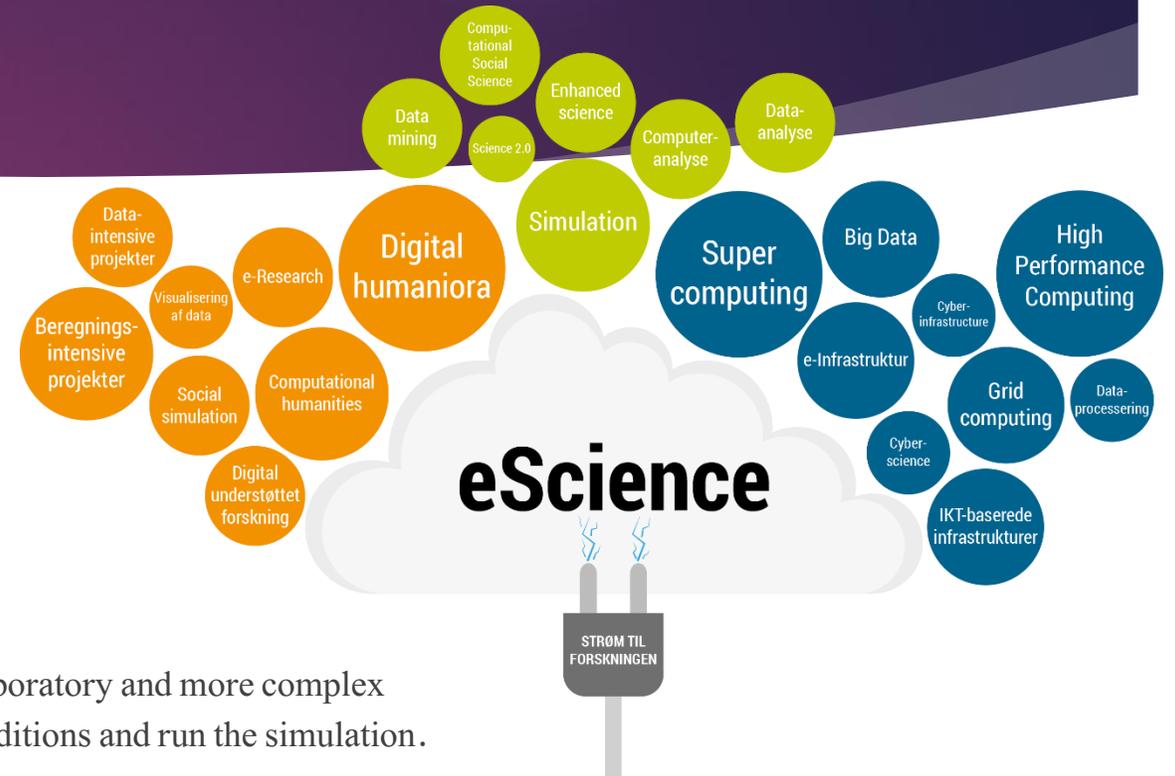
# eScience = Data Science

- ▶ Empirical:
  - ▶ Observe Natural world, Replicate natural world in laboratory
- ▶ Theoretical:
  - ▶ to model the empirical observation use theory
- ▶ Computational:
  - ▶ Simulate in the computer (Problems that couldn't observe in laboratory and more complex that could be Analysis by theoretical models. Define initial conditions and run the simulation.



# eScience = Data Science

- ▶ Empirical:
  - ▶ Observe Natural world, Replicate natural world in laboratory
- ▶ Theoretical:
  - ▶ to model the empirical observation use theory
- ▶ Computational:
  - ▶ Simulate in the computer (Problems that couldn't observe in laboratory and more complex that could be Analysis by theoretical models. Define initial conditions and run the simulation.
- ▶ eScience: acquire massive data sets (databases, visualization, scale out computing, NoSql, machine learning)



# eScience = Data Science

Science is about asking questions

Traditionally: "Query the world"

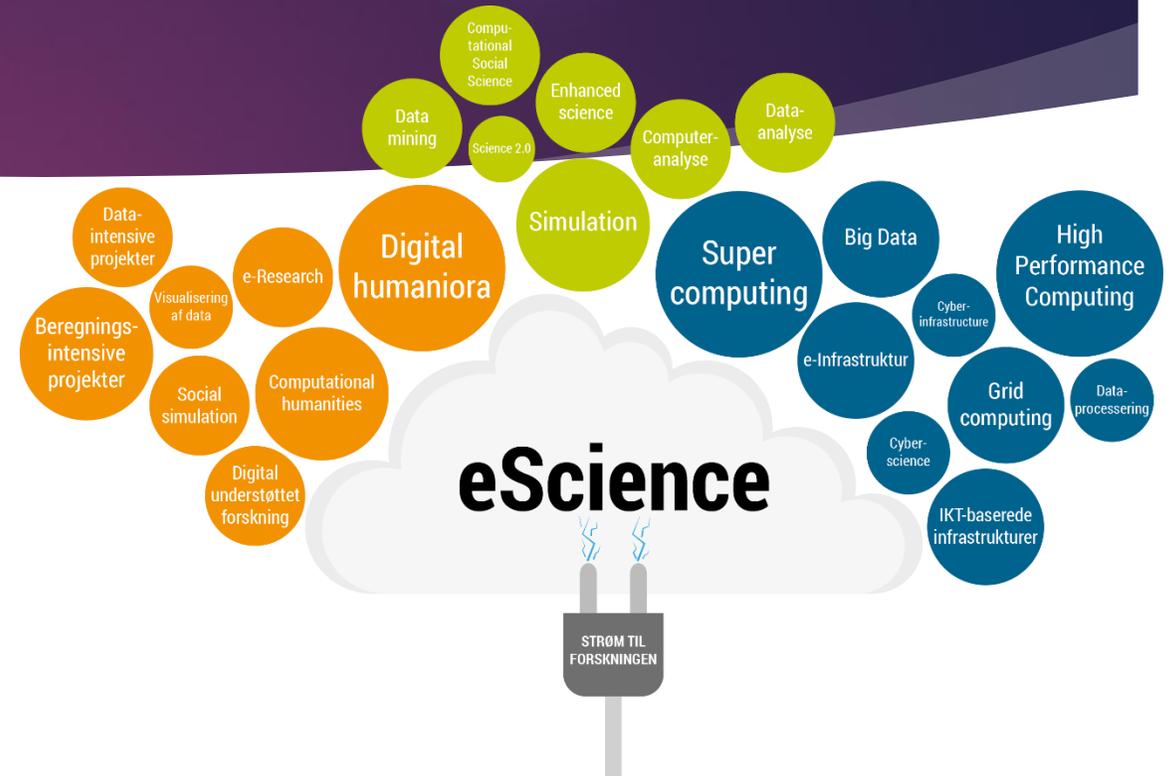
Data acquisitions activities coupled to a specific hypothesis

eScience: "Download the world"

Data acquire in massive in support of many hypothesis

The cost of data acquisition has dropped precipitously

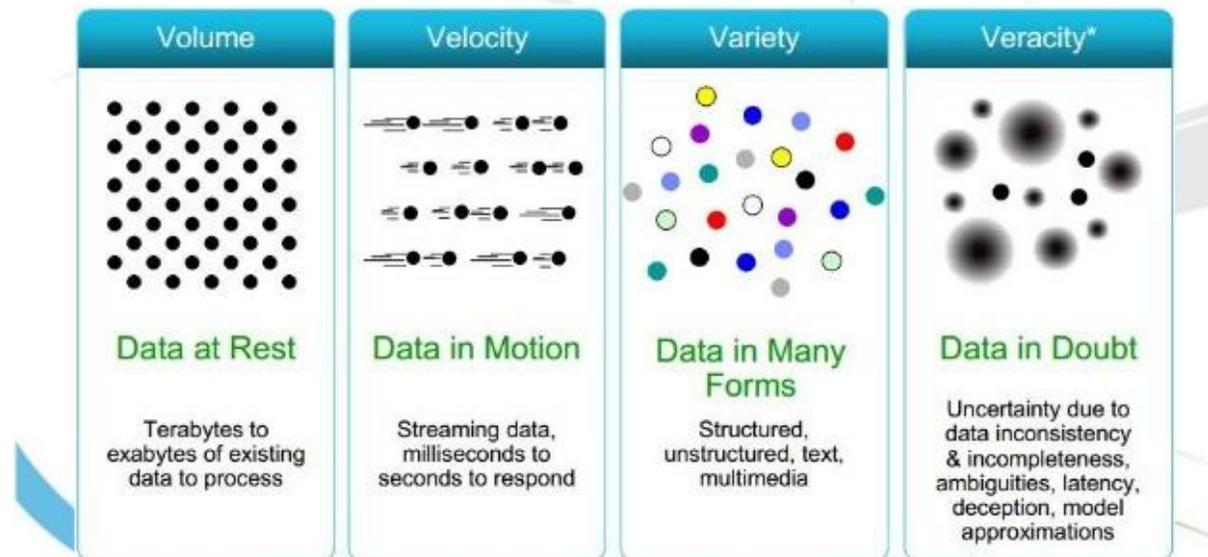
The cost of finding, integrating, analyzing and communicating results is the new bottleneck



<https://vidensportal.deic.dk/what-is-eScience?language=en>

# eScience is about the analysis of data

## 4 V's of Big Data



# Who are data scientists?

- ▶ To be successful, data scientists need an environment that is open, engaging, and fosters collaboration. They need:
  - ▶ Ability to use open source tools they know and love
  - ▶ Enterprise-grade functionality they'll need for critical data science projects
  - ▶ Community that supports them throughout the whole process
- ▶ In this seedbed of innovation, data scientists can break down data barriers and develop ideas that change the world.

# Cloud Providers

- ▶ DataBricks
- ▶ IBM Cloud Data services
- ▶ Google BigQuery
- ▶ DataScience

# DataBricks

Welcome to  databricks™

Community Edition (2.34)

### Featured Notebooks



[Introduction to Apache Spark on Databricks](#)



[Databricks for Data Scientists](#)



[Introduction to Structured Streaming](#)

### New

-  [Notebook](#)
-  [Job](#)
-  [Cluster](#)
-  [Table](#)
-  [Library](#)

### Documentation

-  [Databricks Guide](#)
-  [Python, R, Scala, SQL](#)
-  [Importing Data](#)

### Open Recent

-  [Introduction to Apache Spark on Databricks](#)
-  [Databricks for Data Scientists](#)
-  [cs110\\_lab1\\_power\\_plant\\_ml\\_pipeline](#)
-  [Introduction to Structured Streaming](#)

### What's new?

- Jobs support concurrent runs
- GPU instance types are available
- Github commit messages support international characters

[Latest release notes](#)

[Send Feedback](#)

# A Gentle Introduction to Apache Spark on Databricks

- ▶ **Workspaces**
- ▶ Notebooks
  - ▶ Dashboard
  - ▶ Jobs
- ▶ Libraries (different languages)
- ▶ Tables (Amazon s3)
- ▶ Clusters (groups of computers)
- ▶ Apps (Third party applications, Tableau)

# Spark

- ▶ sparkContext (Apache Spark engine) and SQLContext (DataFrame Functionality)
  - ▶ Spark 2.0 : SparkSession
- ▶ Data Interface :
  - ▶ Dataset
  - ▶ Dataframe
  - ▶ RDD (Resilient Distributed Dataset)

# IBM Cloud Data Service

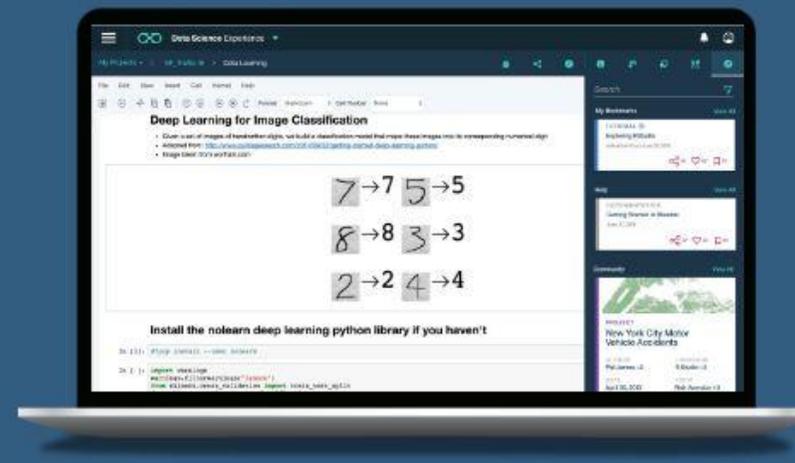
Data Science Experience

## IBM Data Science Experience

Analyze data using RStudio, Jupyter, and Python in a configured, collaborative environment that includes IBM value-adds, such as managed Spark.

Free 30-day trial

View pricing



# Articles + Data sets + Notebooks + Tutorials

The screenshot displays the IBM Analytics Community website interface. At the top, the browser address bar shows the URL <https://apsportal.ibm.com/analytics>. Below the address bar is a navigation bar with the IBM logo and the text "Data Science Experience".

The main content area is titled "Community" and features a search bar with the placeholder text "Search". Below the search bar are navigation tabs for "All", "Articles", "Data Sets", "Notebooks", and "Tutorials". The "Data Sets" tab is currently selected.

The search results are displayed in a grid format under the heading "Search Results". Each result card includes a decorative header image, the title of the data set, the author (IBM), the date, and the topic. Each card also has a heart icon and a bookmark icon at the bottom right.

DATA SET	AUTHOR	DATE	TOPIC
GoSales Transactions for Naive Bayes Model	IBM	Dec 08, 2016	Leisure
GoSales Transactions for Logistic Regression...	IBM	Dec 08, 2016	Leisure
World Tourism Data by the World Tourism...	IBM	Nov 07, 2016	Leisure
Employed population by occupation and age	IBM	Nov 07, 2016	Society
United States Demographic Measures:...	IBM	Nov 07, 2016	Society
SETI data for Kepler 1229b	IBM	Nov 07, 2016	Science & Technology
Primary school completion rate % of relevant...	IBM	Nov 06, 2016	Society
Car performance data	IBM	Oct 18, 2016	Transportation
Worldwide County and Region - National...			
United States Demographic Measures: Income			
Country Statistics: Telephones - Fixed Lines			
Dry Bulb Temperature, by country, station...			

# Data Source

▶ Data Service

▶ External

Amazon Redshift

Amazon S3

Apache Hive

Cloudera Impala

dashDB

DB2

Hortonworks HDFS

IBM Infomix

Microsoft Azure

Microsoft SQL

Mysql

Netezza

Oracle

PostgreSQL

SQL Database

# DataScience Cloud

## Begin your data journey.

The DataScience Cloud empowers every step of your data journey, from ingestion to analysis and beyond.



# Datascience Cloud



## Connect

Connect your data with turnkey integrations or custom connectors.



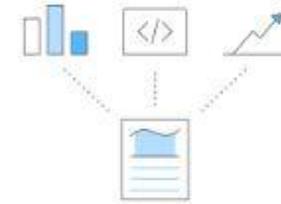
## Explore

Take a deep dive into your data lake with powerful discovery tools.



## Build

Create models and analyses by leveraging your data knowledge.



## Deploy

Instantly deploy, manage, and scale models in production.

# Solutions



DATASCIENCE

[Products](#)[Solutions](#)[Resources](#)[Education](#)[Company](#)[REQUEST DEMO](#)

## Make data science integral to your strategy.

DataScience isn't one-size-fits-all. We provide data solutions that are tailored to your unique needs, no matter what you do.



### C-Suite

Stay competitive with insights and expertise that drive measurable results.



### Marketing

Identify your most valuable customers and channels to optimize marketing spend.



### Product

Deliver better experiences with insight into how customers use your products.



### Finance

Accurately forecast revenue and identify the factors that impact profit the most.



### Data Science

Perform collaborative data exploration and predictive modeling all in one place.



### Customer Support

Improve satisfaction with insight into topics affecting the customer experience.

# Google Cloud Platform for Data Scientists



## GOOGLE CLOUD PLATFORM FOR DATA SCIENTISTS

Scalable, easy-to-use infrastructure and tooling for Data Scientists

[TRY IT FREE](#)

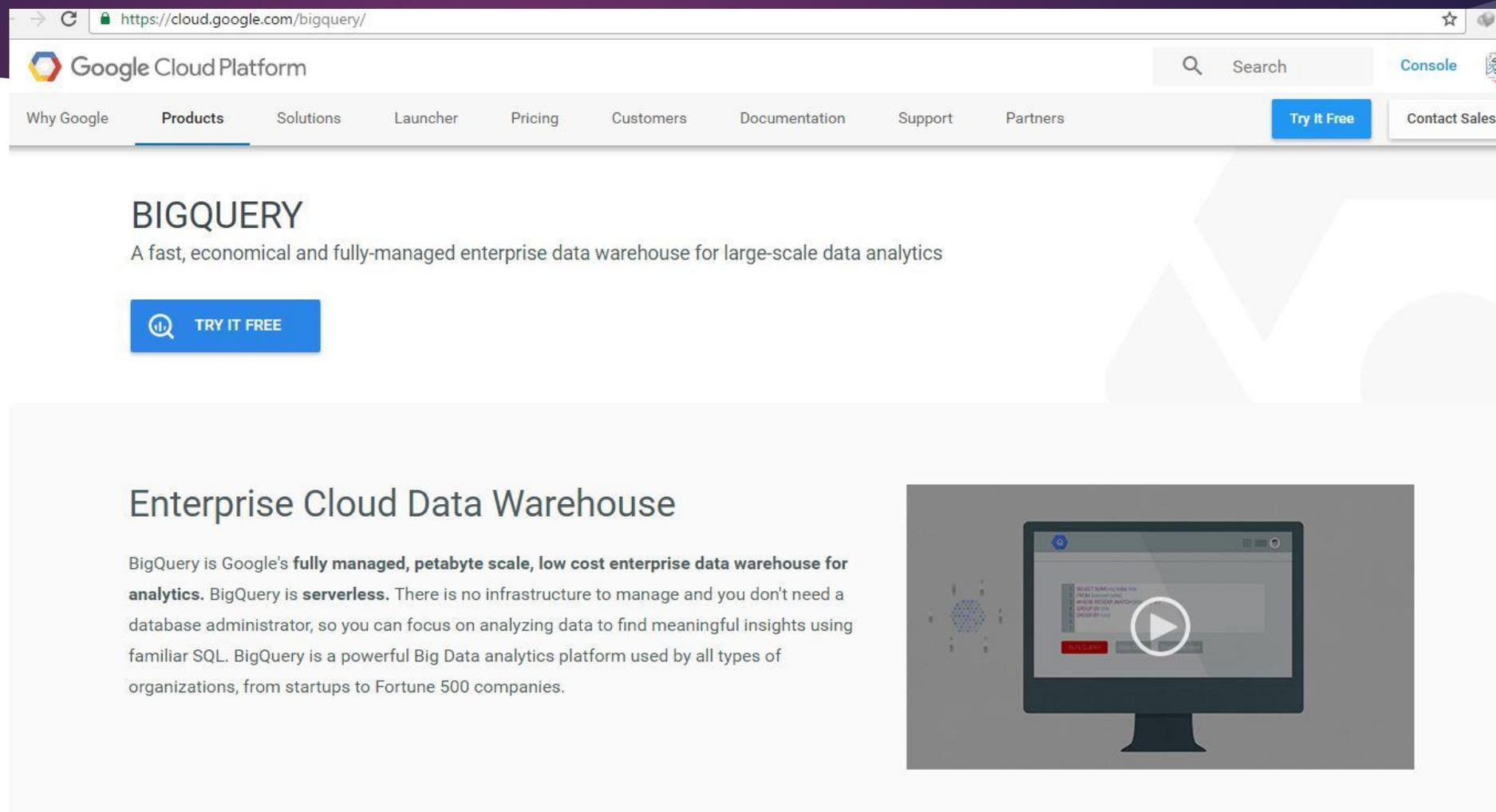
[VIEW MY CONSOLE](#)

## Analyze and Strategize More Intelligently

Google Cloud Platform provides data scientists **key technology and tools to extract tangible business value from massive data assets**. From managed Spark clusters and fast SQL analysis to the latest in machine learning, Google Cloud Platform **empowers data scientists** to spend more time finding value in data and less time worrying about infrastructure. Whether the task at hand is tactical optimization, predictive analytics, nuanced learning, recommendation engines or building automated decision engines, Google Cloud Platform helps Data Scientists **work smarter**.



# Big query (google data warehouse)



→ ↻ https://cloud.google.com/bigquery/

Google Cloud Platform

Search Console

Why Google Products Solutions Launcher Pricing Customers Documentation Support Partners Try It Free Contact Sales

## BIGQUERY

A fast, economical and fully-managed enterprise data warehouse for large-scale data analytics



### Enterprise Cloud Data Warehouse

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



# BigQuery features

- ▶ Speed & Scale : BigQuery can scan TB in seconds and PB in minutes. Stream 100,000 rows per second
- ▶ Incredible Pricing : scale and pay for storage and compute independently (pays-as-you-go model)
- ▶ Security and Reliability : automatically encrypt and replicates your data, fully controlled,
- ▶ Global Availability : store BigQuery data in European locations.
- ▶ Fully Integrated with : SQL, Cloud dataflow, Spark, Hadoop
- ▶ Partnership

The logo for Looker, featuring the word "looker" in a lowercase, sans-serif font with a stylized "o" that has three small circles above it.The logo for Tableau, consisting of a colorful grid of dots followed by the word "tableau" in a lowercase, sans-serif font.The logo for Qlik, featuring the word "Qlik" in a bold, sans-serif font followed by a green circular icon with a white "Q" inside.The logo for Talend, featuring the word "talend" in a lowercase, sans-serif font with a green asterisk at the end.

Google Analytics 360 Suite

The logo for SnapLogic, featuring the word "snapLogic" in a lowercase, sans-serif font with a green "s" and "n" and a purple "Logic".

# Q & A

[Sg.sharif.ir](http://Sg.sharif.ir)



[Telegram.me/BigDataWorkGroup](https://t.me/BigDataWorkGroup)