



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

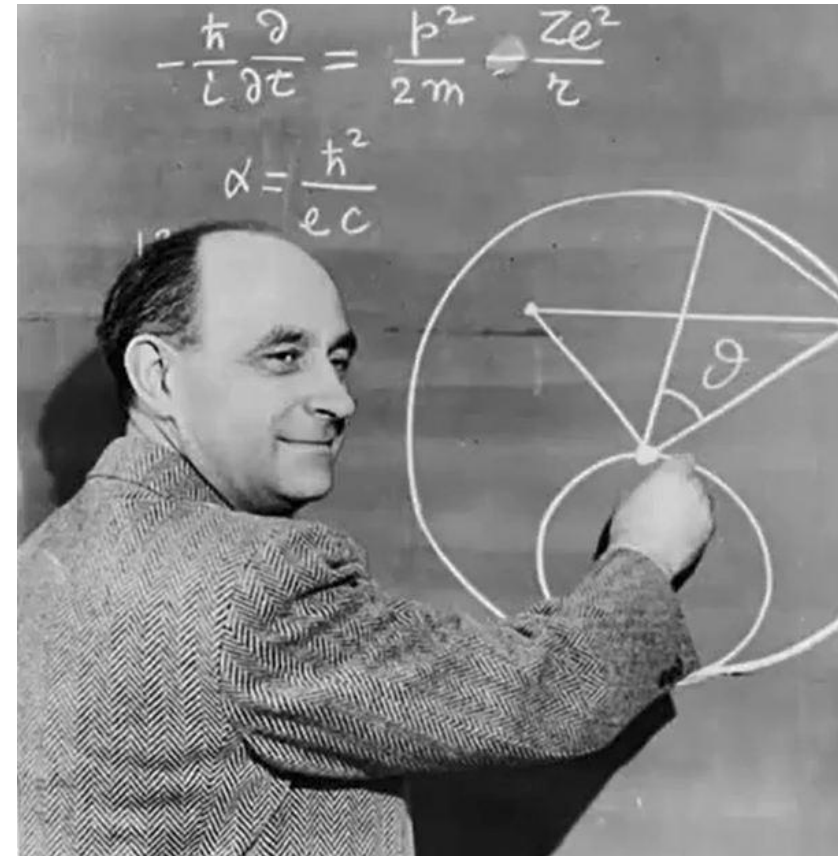
- ▶ Empirical:
 - ▶ Observe Natural world, Replicate natural world in laboratory



public domain

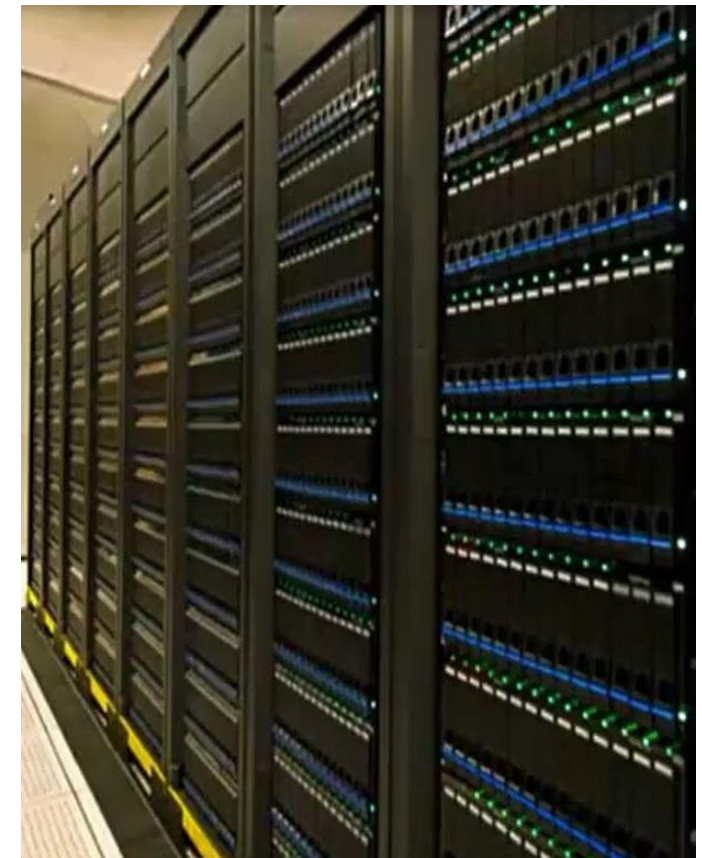
eScience = Data Science

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



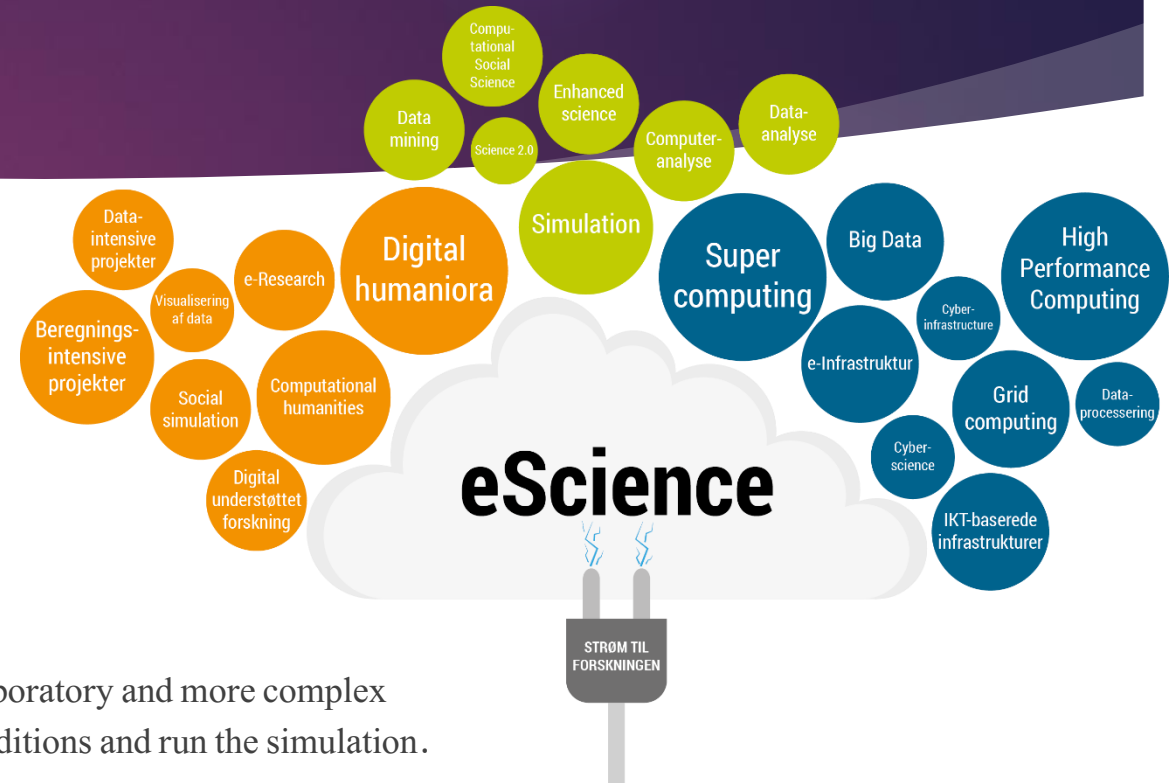
eScience = Data Science

- ▶ 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)



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

eScience = Data Science

Science is about asking questions

Traditionally : “Query the world”

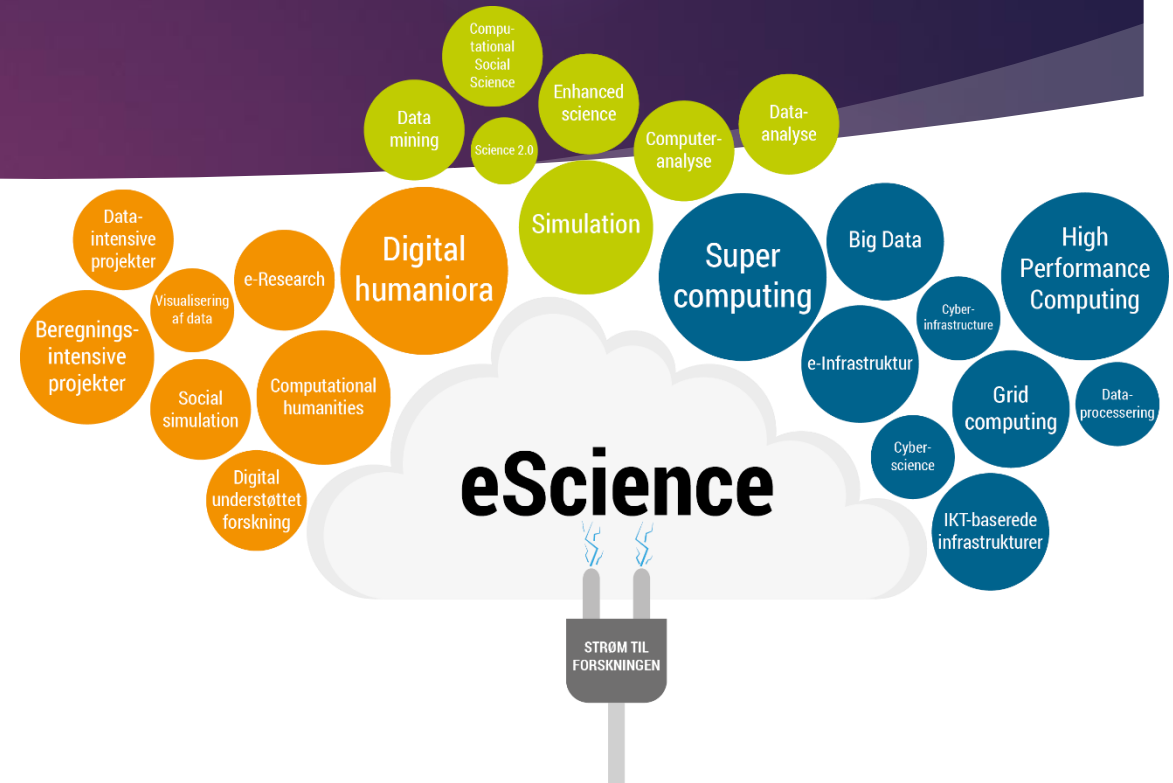
Data acquisitions activities coupled to a specific hypothesis

eScience : “Download the world”

Data acquire in massive in support of many hypothesis

The cost of data acquisition has dropped precipitously

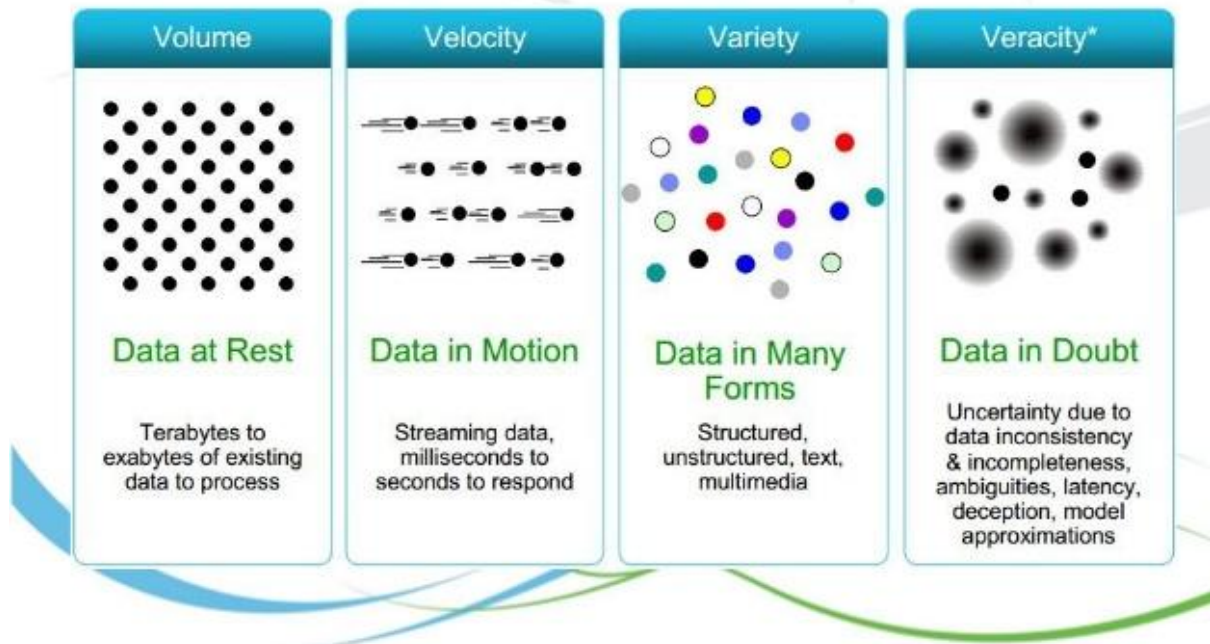
The cost of finding, integrating, analyzing and communicating results is the new bottleneck



<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 scientis?

- ▶ 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

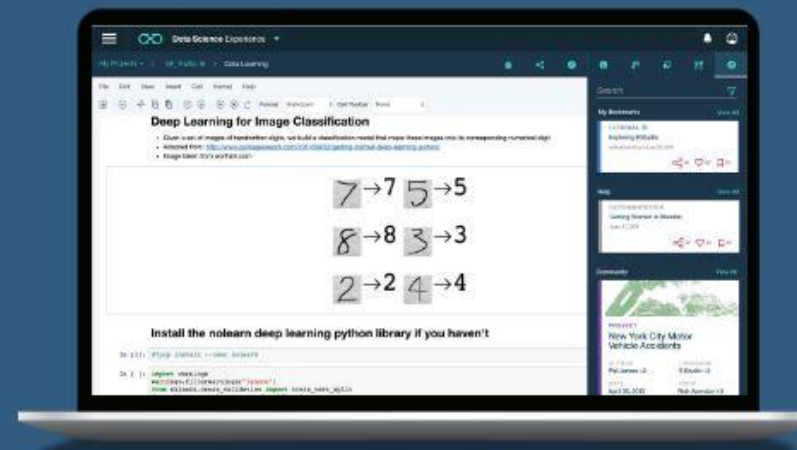
Data Science Experience

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 Portal interface. At the top, a 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 tabs for "All", "Articles", "Data Sets", "Notebooks", and "Tutorials". The "Data Sets" tab is currently selected. The search results are displayed in a grid of 12 cards, each representing a data set. Each card includes a title, author (IBM), date, topic, and a heart icon indicating the number of likes. The data sets shown are:

- GoSales Transactions for Naive Bayes Model (Dec 08, 2016, Topic: Leisure, 0 likes)
- GoSales Transactions for Logistic Regression... (Dec 08, 2016, Topic: Leisure, 1 like)
- World Tourism Data by the World Tourism... (Nov 07, 2016, Topic: Leisure, 1 like)
- Employed population by occupation and age (Nov 07, 2016, Topic: Society, 3 likes)
- United States Demographic Measures:... (Nov 07, 2016, Topic: Society, 5 likes)
- SETI data for Kepler 1229b (Nov 07, 2016, Topic: Science & Technology, 7 likes)
- Primary school completion rate: % of relevant... (Nov 06, 2016, Topic: Society, 4 likes)
- Car performance data (Oct 18, 2016, Topic: Transportation, 17 likes)
- Worldwide County and Region - National... (Topic: Society, 5 likes)
- United States Demographic Measures: Income (Topic: Society, 7 likes)
- Country Statistics: Telephones - Fixed Lines (Topic: Society, 4 likes)
- Dry Bulb Temperature, by country, station... (Topic: Society, 17 likes)

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



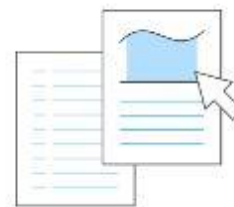
Connect

Connect your data with turnkey integrations or custom connectors.



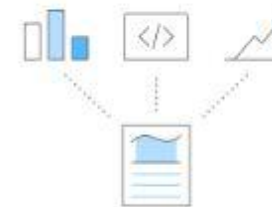
Explore

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](#)

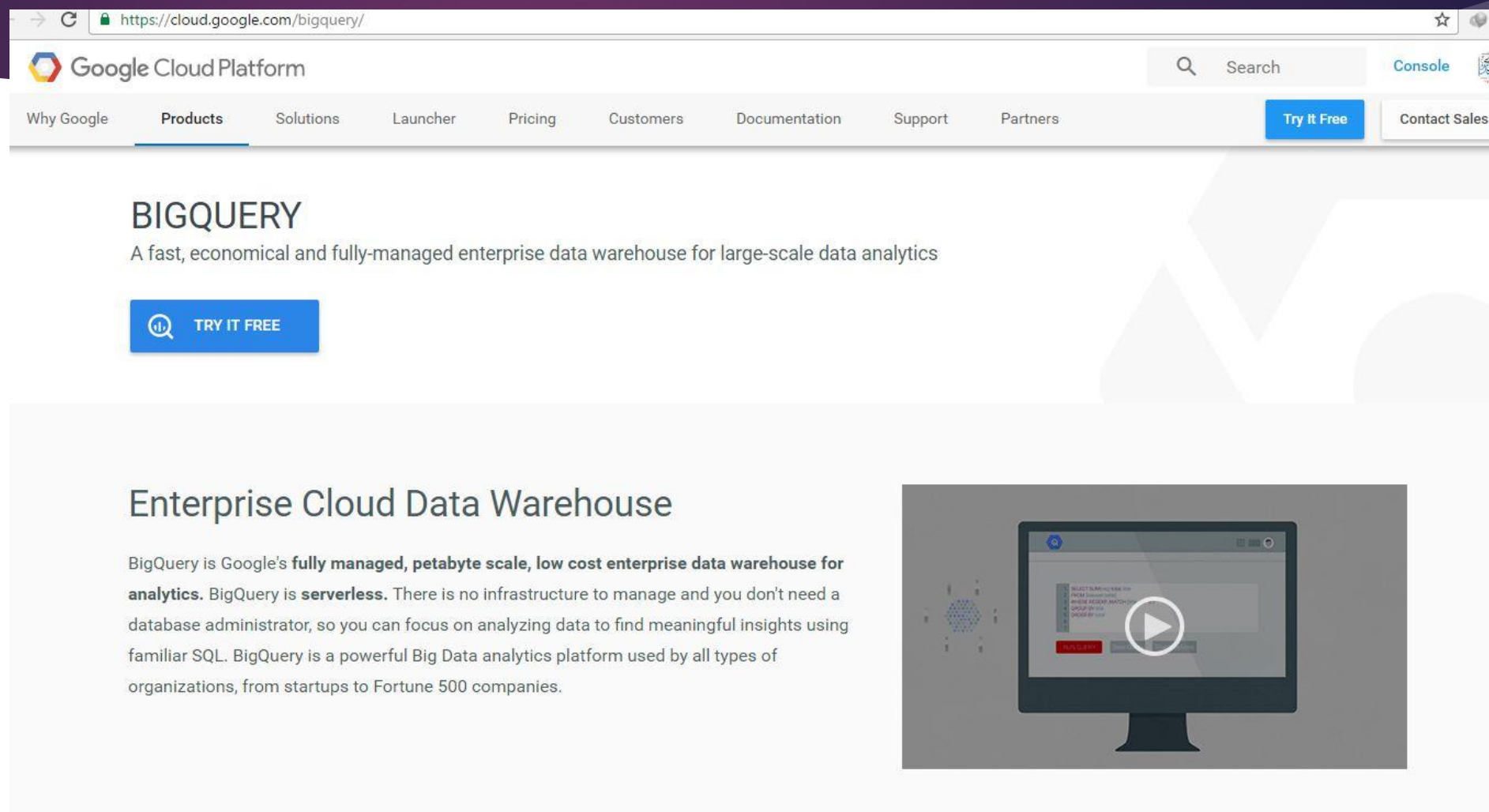
[VIEW MY CONSOLE](#)

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)



The image is a screenshot of the Google Cloud Platform BigQuery landing page. At the top, there's a navigation bar with the Google Cloud Platform logo, a search bar, and links to 'Console', 'Try It Free', and 'Contact Sales'. Below the navigation bar, the main heading is 'BIGQUERY' in a large, bold, sans-serif font. Underneath this heading is a sub-headline: 'A fast, economical and fully-managed enterprise data warehouse for large-scale data analytics'. Below the sub-headline is a blue button with a magnifying glass icon and the text 'TRY IT FREE'. Further down, there's a section titled 'Enterprise Cloud Data Warehouse' in a bold, sans-serif font. Below this title is a paragraph of text: '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.' To the right of this text is a graphic of a computer monitor displaying a SQL query in a code editor. The query is: 'SELECT SUM(sales) FROM [dataset].[table] WHERE YEAR(sales) = 2014 GROUP BY state ORDER BY state'. A large play button icon is overlaid on the monitor screen.

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

TRY IT FREE

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



Google Analytics 360 Suite



Q & A

Sg.sharif.ir



Telegram.me/BigDataWorkGroup