BigDataWorkGroup

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

https://www.coursera.org/specializations/data-science

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

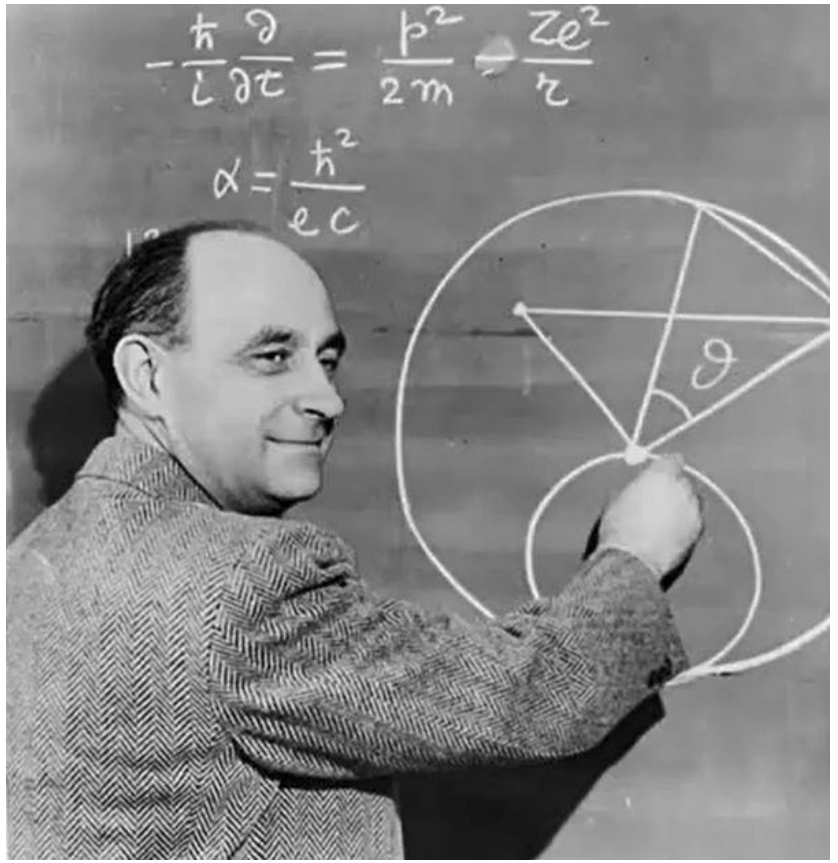
- Empirical :
  - Observe Natural world, Replicate natural world in laboratory

public domain

https://www.coursera.org/specializations/data-science

eScience = Data Science

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



https://www.coursera.org/specializations/data-science

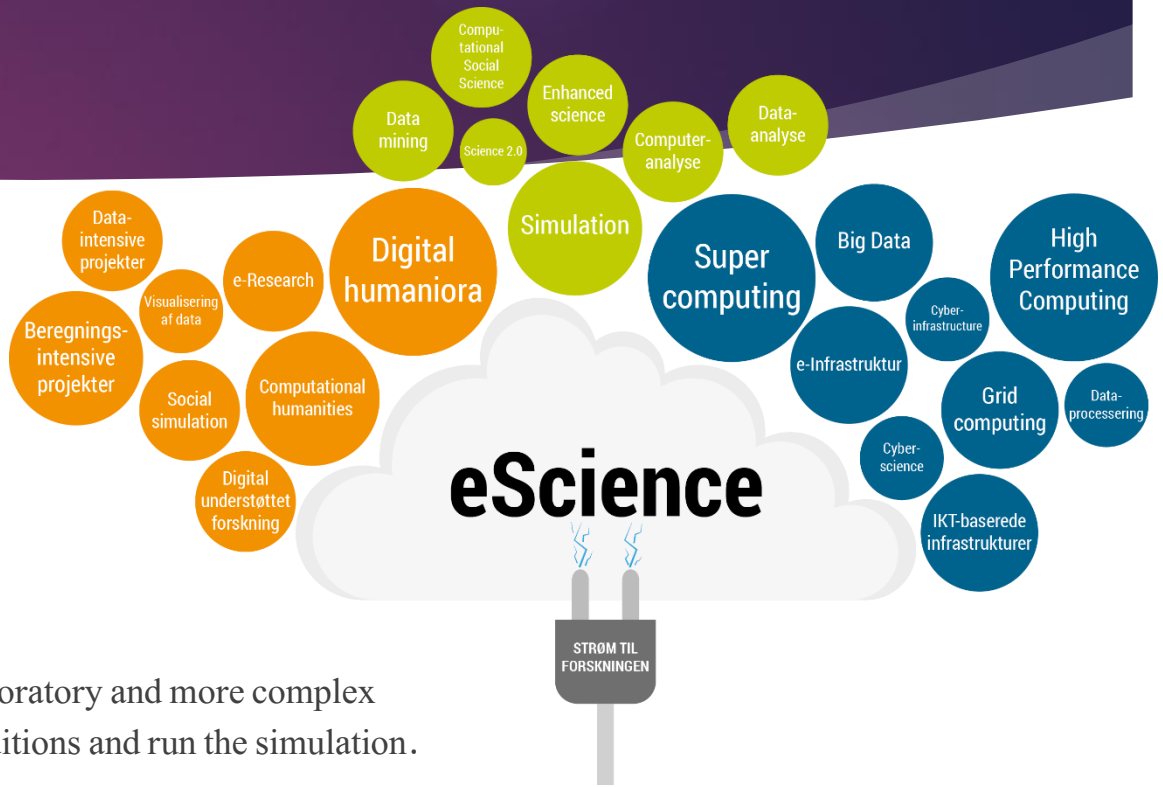
eScience = Data Science

- 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.

https://www.coursera.org/specializations/data-science

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,macinue learing



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

https://www.coursera.org/specializations/data-science

eScience = Data Science

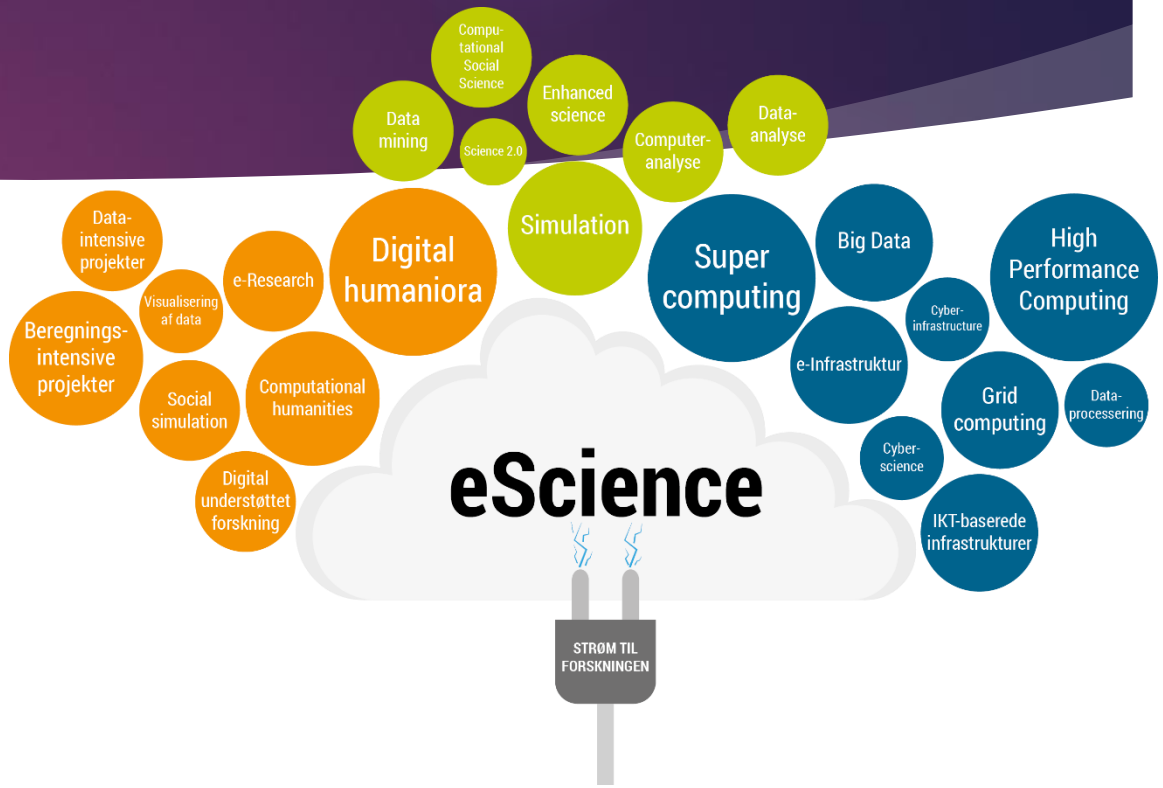
Science is about asking questions

Traditionally : “Query the world”

Data acquisitions activities coupled to a specific hypothesis

eScicence : “Download the world”

Data acquire in massive in support of many hypothesis



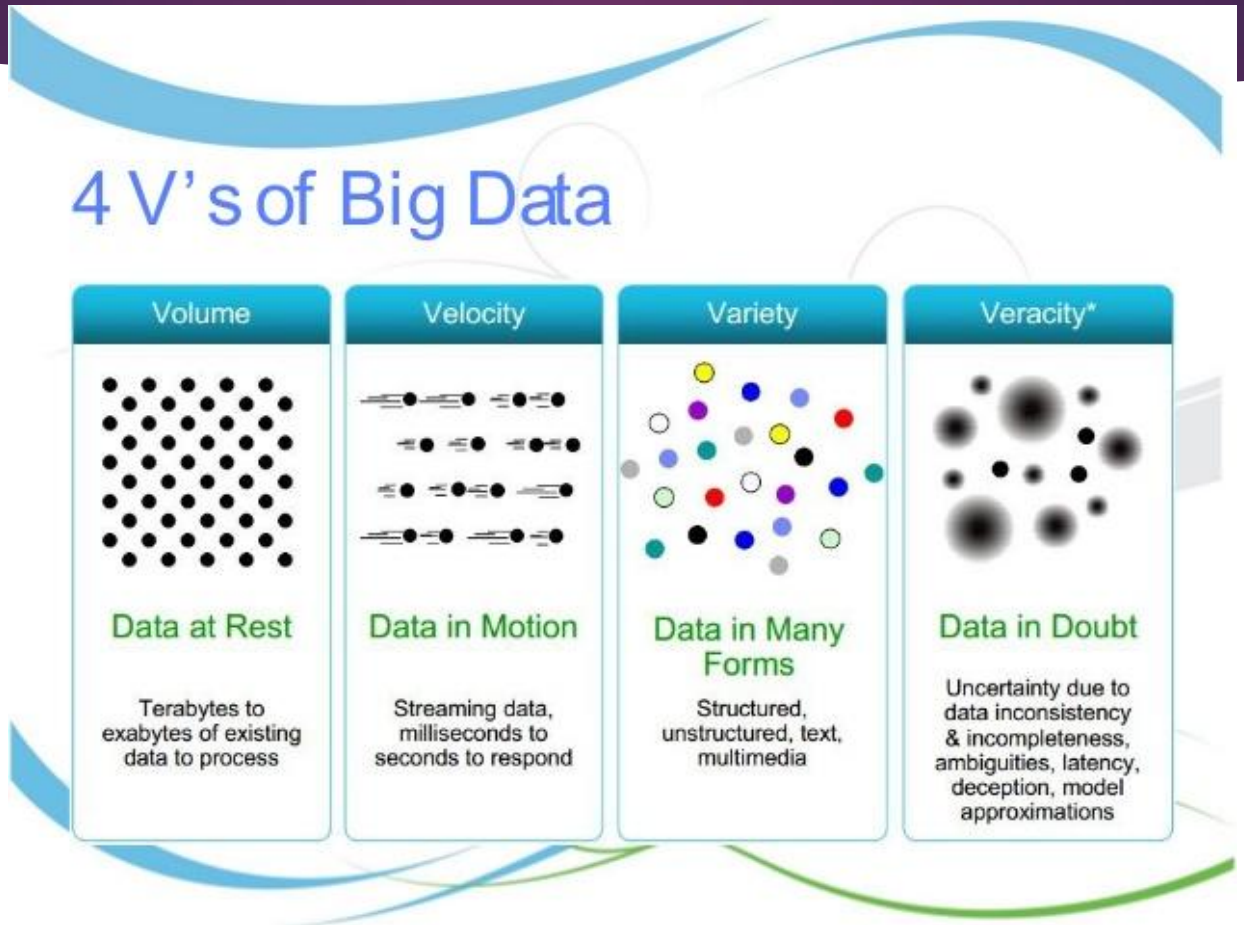
https://vidensportal.deic.dk/what-is-eScience?language=en

The cost of data acquisition has dropped precipitously

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

https://www.coursera.org/specializations/data-science

eScience is about the analysis of data



http://www.slideshare.net/RenuSuren/big-data-analysis-for-page-ranking-using-mapreduce

Who are data scienctis?

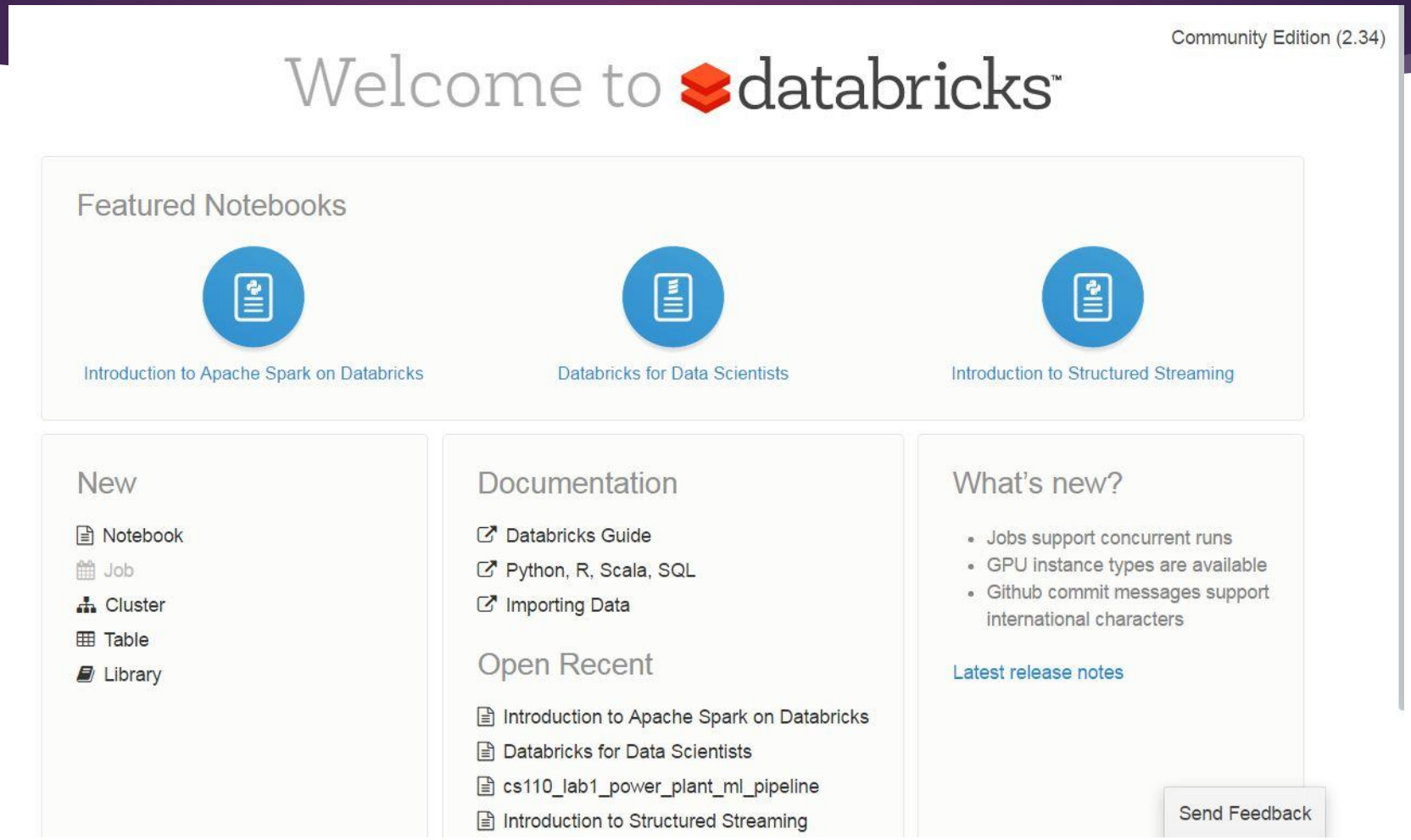
- To be successful, data scientists need an environment that is open, engaging, and fosters collaboration. They need:
  - Ability to use open source tools they know and love
  - Enterprise-grade functionality they'll need for critical data science projects
  - Community that supports them throughout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

http://www.ibm.com/analytics/us/en/technology/data-science/

Cloud Providers

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

DataBricks


Community Edition (2.34)
Welcome to databricks
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

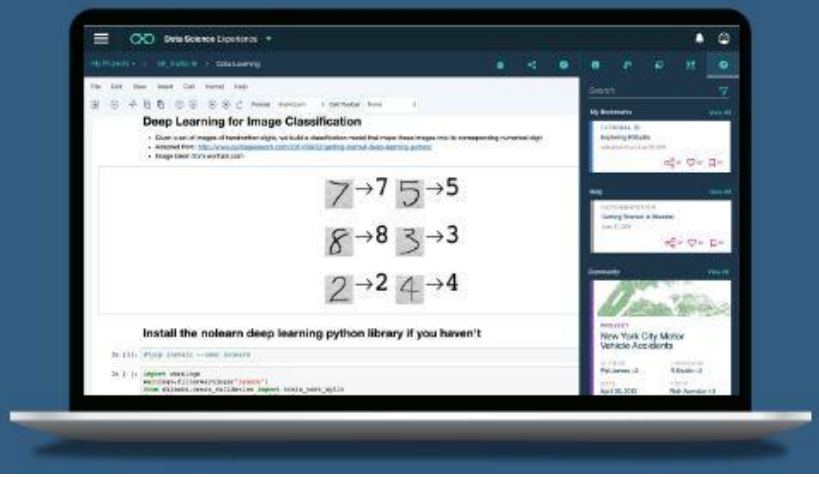
Data Science Experience

IBM Data Science Experience

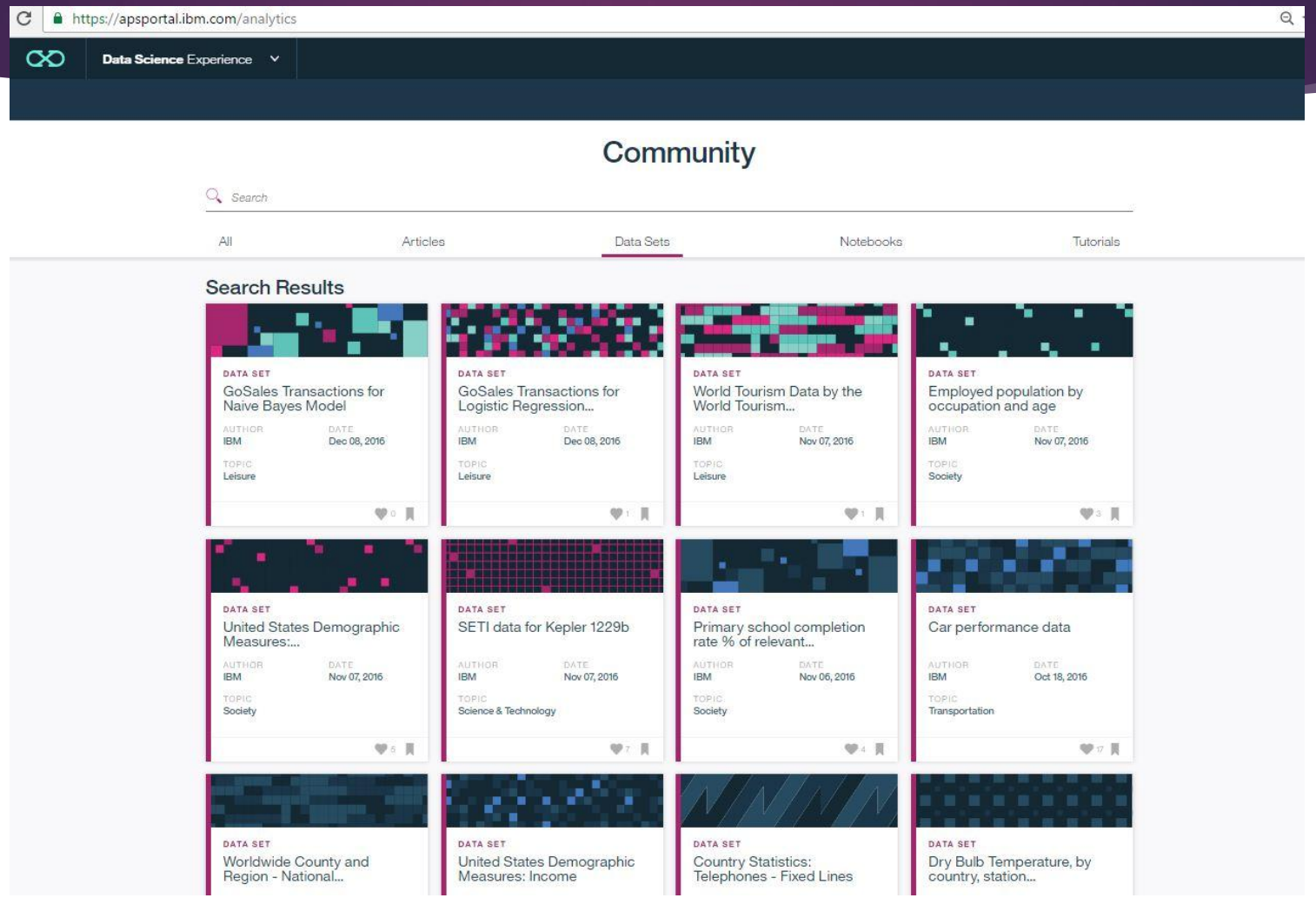
Analyze data using RStudio, Jupyter, and Python in a configured, collaborative environment that includes IBM value-adds, such as managed Spark.

Free 30-day trial

View pricing



Articles + Data sets + Notebooks + Tutorials



Data Source

- Data Service
- External

Amazon Redshift

Amazon S3

Apache Hive

Cloudera Impala

dashDB

DB2

Hortonworks HDFS

IBM Infomix

Microsoft Azure

Mircosoft SQL

Mysql

Netezza

Oracle

PostgreSQL

SQL Database

DataScience Cloud

Datascoence Cloud

Connect

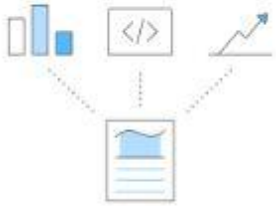
Connect your data with turnkey integrations or custom connectors.

Explore

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

Build

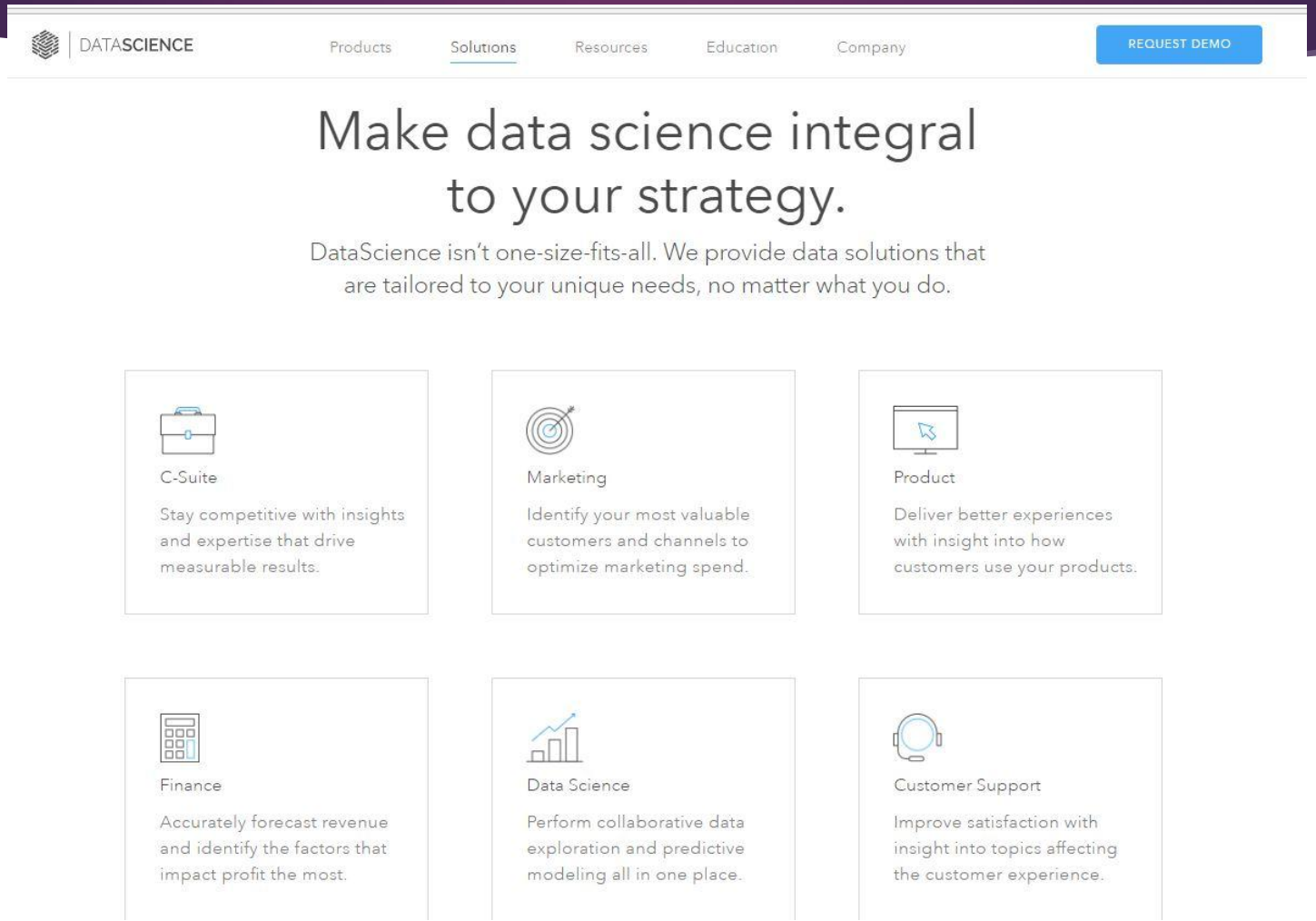
Create models and analyses by leveraging your data knowledge.



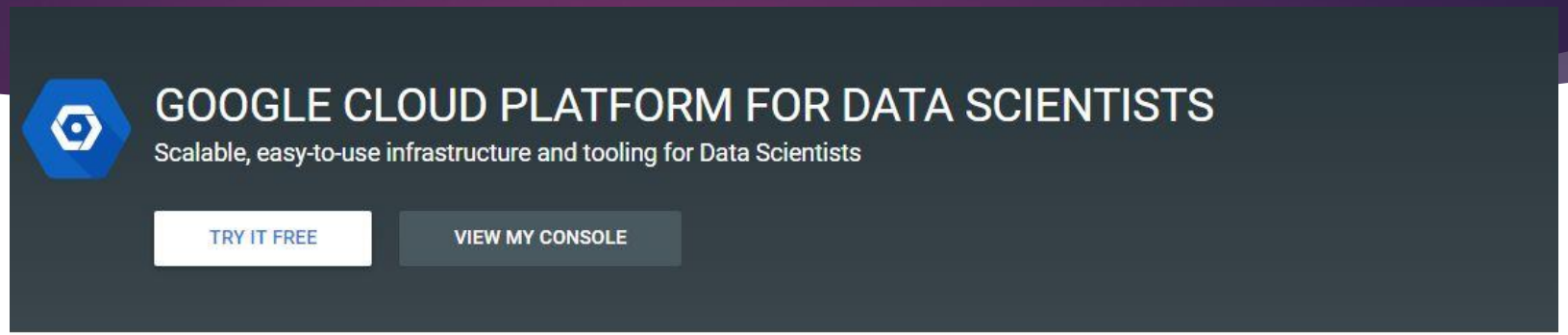
Deploy

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



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

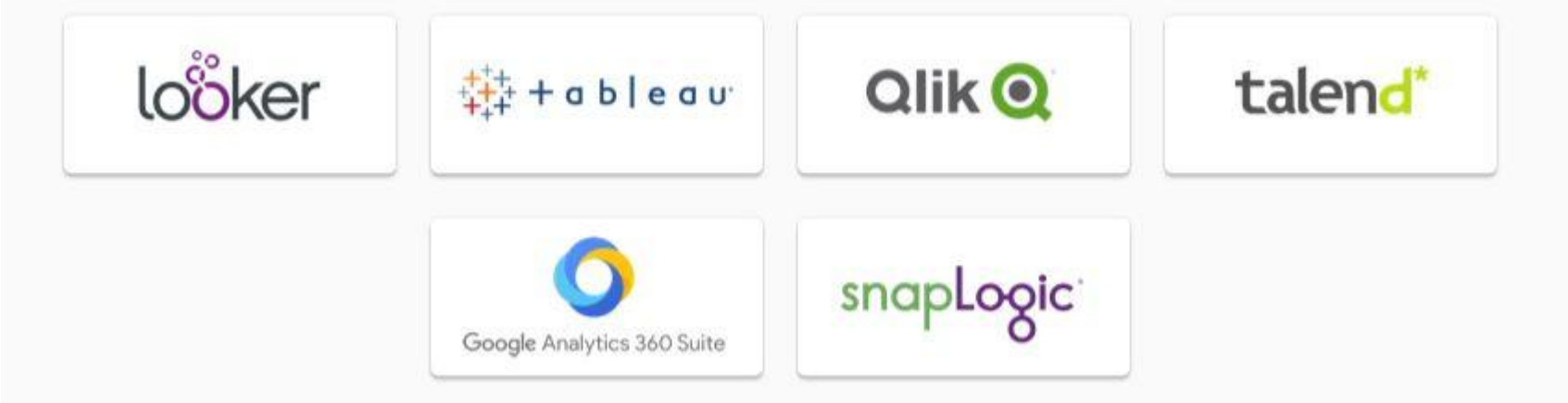
TRY IT FREE

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



Q & A

Sg.sharif.ir

Telegram.me/BigDataWorkGroup