Apache Spark:
The New Language Of Analytics

Lightning-fast cluster computing

Presented By:

Wise With Data

Manage With Confidence

In Partnership With:

MARINER INNOVATIONS

About WiseWithData

- Founded in 2015, WiseWithData is Canada's leading data science consulting firm focused on scalable open-source analytics technology
  - ➔ Management Consulting
  - ➔ Technical Consulting
  - ➔ Analytical Products & Solutions
  - ➔ Education

- We partner with leading technology firms like Mariner Innovations
- Some of our recent clients include:

Deloitte.
MD Financial Management CMA Companies
Third Bridge
Health Canada
Immigration, Refugees and Citizenship Canada
Employment and Social Development Canada
Indigenous and Northern Affairs Canada
WorkSafeBC
MindBridge Analytics Inc.
House of Commons
Chambre des communes
Service Canada

Solving Problems With Data

Horizontal Solutions

- BI/Data Visualization
- Data Warehousing
- Forecasting
- Fraud & Security Intelligence
- IT Service Management
- Marketing Optimization
- Predictive Maintenance
- Risk Management

Vertical Solutions

- Finance / Banking
- Financial Accounting & Audit
- Government
- Healthcare
- Insurance
- Logistics
- Retail
- Telecom

Data Science In A Nutshell

Descriptive Analytics
- BI and Reporting
- Data Visualization
- Advanced Search
- Graph Processing
- Natural Language Processing

Predictive Analytics
- Predictive Models
- Forecasting
- Classification
- Clustering
- Recommendation

Prescriptive Analytics
- Test & Learn Strategies
- Simulations
- Decision Analysis
- Lifetime Valuations / NPV models
- Operations Research / Optimization

Data Governance
- Data Security and Ownership
- Data Privacy
- Master Data Management
- Metadata Management
- Data Monitoring

Data Management
- Data Profiling
- Data Quality & Cleansing
- Data Matching
- Data Integration / Transformation
- Data Warehousing

Why It's Time To Ditch "Big Data"

- The term is no longer relevant
- It's now synonymous with analytics in general, which denigrates the enormous value you can get from "Small Data"
- Data size should not dictate your choice of analytics software or your approach to solving a business problem

And Start Talking About

Analytics Software Tools That Are...

- Open and Free
- Fast and Efficient
- Simple To Use
- Scalable and Fault-tolerent
- Comprehensive In Capabilities

Data Science Platforms

An Ideal Data Science Platform Would Provide

- A full set of capabilities for all analytics uses
- Little need to copy or duplicate data
- Integrated workflows within the same tool and seamless of integration of multiple tools
- The ability to quickly iterate

A Brief History Of Data Science Platforms

Expensive – Proprietary
Slow – Unscalable

Inexpensive – Open-Source
Fast – Scalable

- 1968/1969+ IBM Cognos, SPSS, etc.
- 1976 - SAS
- 2012/13 Hadoop 1/2
- 2014 - Spark 1.0
- 2016 - Spark 2.0

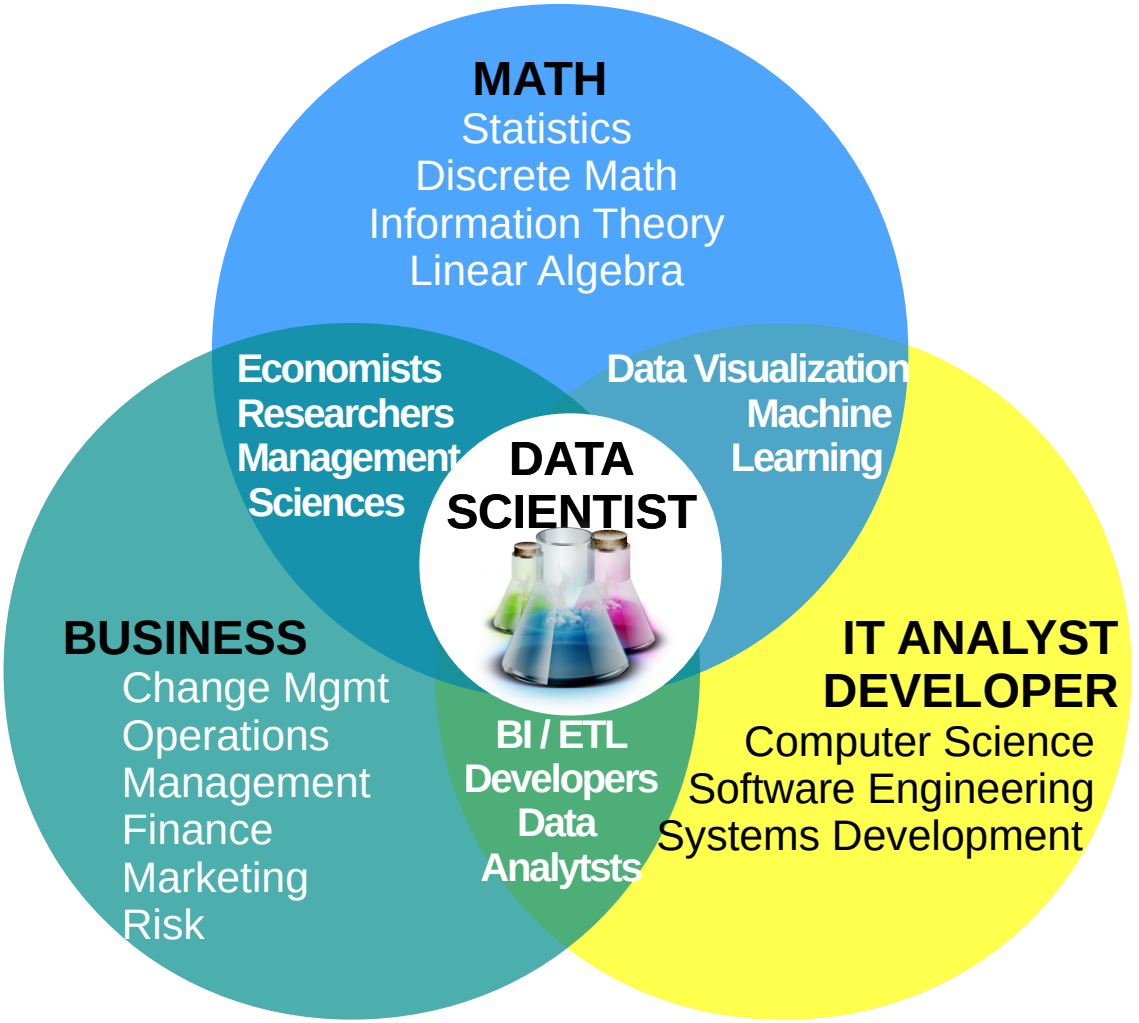
IBM: A large proprietary collection of over 17 vendor products purchased by IBM between 2005 and 2011. Core software designed in 1960's (SPSS-1969, Cognos-1968).

SAS: A proprietary data science platform. It contains a wide variety of analytic tools and methods, mostly developed in-house.

Haddop: A scalable open-source software platform for scalable, distributed computing. Through a set of ecosystem tools, Hadoop can provide analysis of both structured data and unstructured data.

Apache Spark: A open-source high-performance scalable data processing engine. It handles structured & unstructured data, and contains advanced analytical & graph capabilities. An ecosystem of associated tools provide visualization.

Why Data Scientists Are So Rare



Data Science Professionals Have The Need For Speed

Business Problem

Solution

80%

20%

Data preparation is 80% of the effort

Solving the problem is 20% of the effort

<u>The Need For Speed</u>
Simple, Powerful API's + Fast Execution = Rapid Iterations
Rapid Iterations = Agile High Quality Solutions

Wise With Data — Manage With Confidence

Spark Is A Data Scientist's Dream

Invented by Canadian computer scientist Matei Zaharia,
Apache Spark's a revolutionary data science & analytics engine

- **100's of times faster** and more efficient then legacy systems
- **Fully scalable and fault tolerant**, it runs fast and efficiently on everything from laptops to massive clusters of 1,000's of servers
- **Simple easy-to-use coding interfaces** (API) including SQL which can handle streaming or batch applications in the same code
- **Built-in libraries** for data access, streaming, data integration, graph processing, and advanced analytics
- **2nd Most active open-source project (after Linux)** - commercially supported by WiseWithData, Databricks, Hortonworks & Others

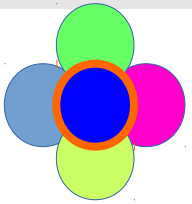
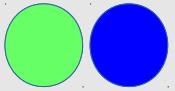
Disruptive Spark Adoption

Web/Retail | Industrial/ High-Tech | Financial | Media/ Entertainment | Life Sciences | Utilities/NGO/ Government

Why?

- ✔ **Rapid Low Cost Deployment**
- ✔ **Powerful Yet Simple Interfaces**
- ✔ **Exceptional Speed & Productivity**

Why Is Spark So Disruptive?

	Proprietary / Legacy	Spark	Impact
Hardware	**$$$$$$$$$$$$**	**$**	• Faster TTV* • Improved ROI
Software Licence	**$$$$$$$$$$$$$$$$$$$$** **$$$$$$$$$$$$$$$$$$$$**	Free + Optional Support (**$$$**)	• Faster TTV* • Flexibility
Source Code	**Closed**	**Open**	• Security • Flexibility • No vendor lock-in
Tools / Applications			• No Integration Mess • Faster TTV*
Execution Speed	**1X**	**100X+**	• Faster TTV* • Faster & Better Results
Productivity	**1X**	**5X+**	• Better Results • Lower HR costs
Streaming / Batch			• Re-usability • Flexibility

*TTV=Time To Value

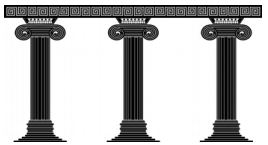
Time To See Spark In Action

The Secrets Of Spark's Success

- **Be Fault Tolerant, Not Wasteful** – Replicate lineage, not work

- **Use Your Memory** – Reduce storage I/O usage by leveraging memory

- **Be Lazy** – Use a lazy execution model to plan and optimize work in advance

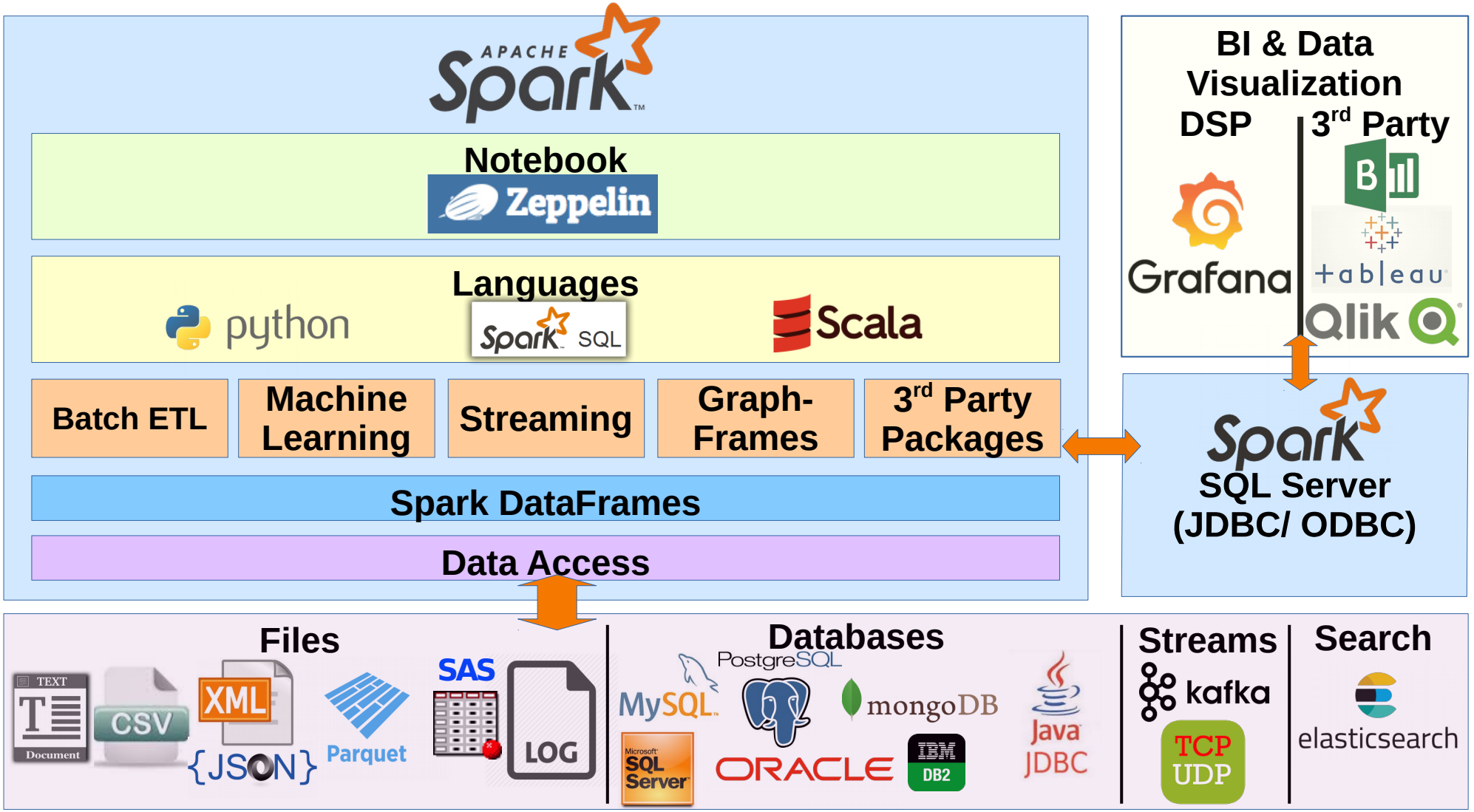


- **Go Columnar!** - Use partitioned columnar data structures like Parquet

- **Keep It Simple Stupid (KISS)** – DataFrames are powerful, simple, and easy to learn

- **Speak Your Language** – You can choose between Python, SQL, Scala, R, or Java

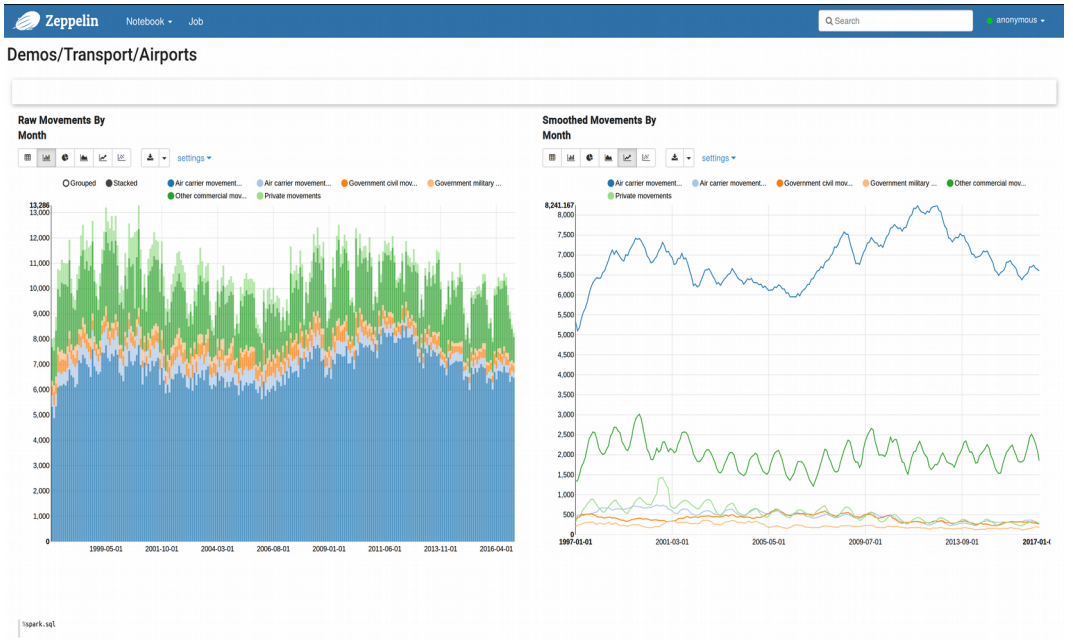
Making Spark Simple To Deploy & Use WWD Data Science Platform (DSP)



DSP Spotlight - Apache Zeppelin Simple & Powerful Spark Notebook

Apache Zeppelin

- A multi-purpose Notebook interface runs on top of Spark for data science use cases such as:
  - Documentation
  - Data Ingestion
  - Data Discovery
  - Advanced Analytics
  - Data Visualization
- Makes using Spark simple and painless
- Create simple dynamic forms for beautiful analytic applications



DSP Is The Foundation Of Our Leading Tools & Solutions

Industry Solutions

GuardDog

Full featured security intelligence suite. It includes business rules, anomaly detection, predictive models, network analysis and investigative capabilities for:

- Fraud, Waste and Abuse
- Public Safety and Border Security
- Anti-Money Laundering
- IT Operations Management

LogHaus

Log ingestion and data warehousing solution. It efficiently transforms log files or streams into highly structured data for reporting. Supported log formats include:

- Firewall Logs
- Web Application Logs
- Windows Event Logs
- Generic SYSLOG formatted logs

Analytic Tools

MatchBox

Comprehensive data quality tool for matching, standardization and entity resolution. It can process address, personal identity and business name data.

PreViso

Business forecasting tool for developing large scale hierarchical forecast models. It includes model reconciliation and monitoring capabilities.

ModelEyes

Tool for model monitoring and governance. Provides standardized reports for assessing the stability and validity production models.

Foundation

WWD Data Science Platform

Data Management & Governance	Data Integration & Transformation	Data Visualization & Reporting	Advanced Analytics

Fast, Simple, Scalable Analytics
Your Spark Journey Starts Today

WiseWithData & Mariner Innovations

- Spark based analytical products and solutions
- Spark advisory, deployment, development, education, and support services
- WiseWithData.com / MarinerInnovations.com
- Contact Us:
  - Ian Ghent (Ian.Ghent@WiseWithData.com)
  - Karen Fegarty (Karen.fegarty@marinerinnovations.com)

Spark+AI Summits

- San Francisco,USA - June 4-6, 2018
- London, UK – October 2-4, 2018

Spark Online Resources

- spark.apache.org
- Apache Spark YouTube Channel
- Apache Spark StackOverload list

Wise With Data — Manage With Confidence