The Business Case For Spark, Kafka & Friends

Strata + Hadoop World NY 2015 • Edd Dumbill • @edd
Big Data...

It’s really about business agility
BUYING AGILITY

• Linear scale-out cost
• Opex vs capex
• Ease of purchase
Different use cases put different demands on the data infrastructure.

Increasing cost per unit of capability from scale-up architectures causes rationing of resources. Only the most valuable use cases are pursued.
Scale-out systems move us from managing scarcity to promoting utility
• Architectural factors
  • Schema on read
  • Rapid deployment
  • Mirror production setup
  • Executes faster
• Programmer factors
  • Fun to program
  • Concision
  • Easier to test
  • Faster to write
DATA PLATFORM

Data Management
- Security, Operations, Data Quality, Meta Data Management and Data Lineage

Data Acquisition
- Internal
- External

Low Latency Access
- Data Ingest
- Offline Processing
- Real-Time Processing
- Persistence
- Batch Processing

Data Repository

Analytics

External Systems

Data Services
What is Apache Spark?

- In-memory distributed computing platform
- Comes from Berkeley AMPlab
  - From the same stable: Mesos, Tachyon
- In production with early adopters
- Doesn’t need Hadoop, but runs easily on top
Use cases

- Managing a major retailer's inventory across a diverse network of entities in near real time
- Managing and processing event streams for online gaming
- Supporting data science initiatives across massive data sets at a media analytics company
Why should business care?

• Spark enables use cases Hadoop didn’t provide (streaming, interactive analytics, machine learning, graphs) all in one platform
• Spark is fast
  • Iteration time down, more productive
• Spark can use existing cluster investment
  • Sits on HDFS storage, can run under YARN (but also Amazon S3, or Cassandra)
Why should business care? (2)

- Spark speaks SQL
  - Use SQL skills and tools, e.g. Tableau
  - Spark Dataframes integrates external data sources into one context: RDBMS, Hive, JSON...
- Spark is developer-friendly
  - Concise and fluid to program
  - Language integration: Scala, R, Python, Java
What is Apache Kafka?

• Scale-out fault-tolerant messaging system
• Comes from LinkedIn
• Supported by Confluent
Use cases

- Stream processing
- Log aggregation
- Creating decoupled evented architectures
Why should business care?

• Kafka provides scalability in a critical area of distributed applications where it didn’t exist before
• Kafka provides online reliability, compared to alternatives
• Will progress to be a core building block of distributed data architecture
What is Docker?

- Container technology: bundles every part of an application
- Provides isolation for each application without the overhead of running a virtual machine
  - Ships only the parts that are needed—leaves out the operating system
Why should business care?

• Docker makes better use of server resource than virtual machines
• Docker provides a fast and reliable way of deploying applications: it’s the ideal packaging mechanism for scale-out distributed systems
• Docker makes it easy for developers to work in an environment identical to production
  • Sharing containers leads to innovation
What are Notebooks?

- Interactive documents that contain a program and its output
  - Long history: Mathematica
- Particularly successful with data science
- Projects to watch
  - Jupyter [https://jupyter.org/](https://jupyter.org/)
  - Apache Zeppelin [https://zeppelin.incubator.apache.org/](https://zeppelin.incubator.apache.org/)
Simple spectral analysis

An illustration of the Discrete Fourier Transform

\[
X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn}
\]

using windowing, to reveal the frequency content of a sound signal.

We begin by loading a datatfile using ScPy's audio file support:

```
In [1]: from scipy.io import wavfile
rate, x = wavfile.read('test_mono.wav')
```

And we can easily view its spectral structure using matplotlib's built-in `specgram` routine:

```
In [2]: fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))
ax1.plot(x); ax1.set_title('Raw audio signal')
ax2.specgram(x); ax2.set_title('Spectrogram');
```

![Screenshot from ipython.org](ipynb-viewer.png)
Why should business care?

• Notebooks allow easy collaboration and sharing of data science (think “docker for analysis”)
• Notebooks allow analysts and data scientists easy access to data and compute resource
• Notebooks are a building block for enabling employees with more self-service analytical capabilities
  • Commercial version of this is Databricks Cloud
Data is your business
SILICON VALLEY’S DATA MACHINE

[Logos of various companies such as Uber, Tesla Motors, Stitch Fix, Flatiron, Evolv, and IFTTT]
THE EXPERIMENTAL ENTERPRISE

Data science allows us to observe our experiments and respond to the changing environment.

We need to both support investigative work and build a solid layer for production.

The foundation of the experimental enterprise focuses on making infrastructure readily accessible.
BECOME DATA NATIVE

• Can only win with situational awareness
• New architectures offer new opportunities
• Creation of data-driven value requires new approach
• Create an Experimental Enterprise
• Business must lead, and understand the potential of the technology
Edd Dumbill

edd@svds.com
@edd

Yes, we’re hiring!
info@svds.com