Streams: Transform your business one millisecond at a time

Manny Puentes
Founder & CEO, Rebel AI
Who I am

Manny Puentes
Founder CEO, Rebel AI
@epuentes

20 years leading product and strategy in the advertising and software industries
Rebel AI builds products to secure ad delivery and digital identity in programmatic media trading.

www.rebel.ai
@Rebel_AI
Big Data is Kind of a Big Deal.
The Internet of Things is ushering in an unprecedented era of data creation. The companies that will come out on top will be those that can successfully collect, store, and process a large amount of data within milliseconds—and report on it in real-time.
The Global Internet of Things

26 Smart objects per person by 2020

The IoT world will include cars, wearables, home automation, healthcare, & more
Case Study: Digital Advertising

- In seconds, the ad server has to decide the **right ad** to play for the **right person** at the **right time**... billions of times per day.

- Highly transactional & tied directly to revenue; need to be able to reprocess data quickly & effectively

$66B$ spent in digital advertising in 2016, surpassing TV
This requires a data pipeline that can scale as more & more events come online.
Expect Everything to Fail.
A Data Pipeline That Scales Businesses

**Collect**
No matter what the source, data needs to be managed and streamed.

**Store**
Big data requires a fault-tolerant, replicated space where it can be accessed efficiently.

**Segment**
Data is diverse, and making sure the right data gets to the right place is vital.

**Vizualize**
With the right tools, data can now be presented in new and innovative ways.

- **kafka**
- **MapR**
- **CASCADING**
- **druid**
The Data Pipeline Flow

Event Source → Events Server → kafka

CASCADING → MAPR → Spark → Analytics

OOZIE

druid

Journaling
Don’t see the latest, greatest thing? Just remember...

Cool doesn’t scale.
Here are the tools that have worked for us.
Centralized messaging for all components of a data platform

- Fast and Scalable
- Durable
- Messages are persistent on disk and replicated within the cluster

- Distributed by Design
- Cluster-centric design makes it fault-tolerant
Spark

Engine for large-scale data processing

- Combines SQL, streaming, and complex analytics
- Write applications quickly
- Run programs faster than Hadoop
- Access diverse data sources
- Fault-tolerant streaming
Enterprise deployment of Hadoop provides scalable framework to grow big data businesses

- Half the infrastructure needed to process large amounts of data
- Open Source Ecosystem Integration
- Support – Insurance Policy
- Developers don’t have to solve big data problems
- Built-in NFS reduces reliance on Java programmers; instead rely on standard IT resources
Cascading

Robust, reliable, and data-oriented application development platform for Hadoop

- Build & test locally — deploy at scale in production
- Eliminate compute fabric lock-in
- Reduced development & operational complexity
- Data processing code/algorithms
- Runs on & can be ported between MapReduce, Tez, Flink
- Cascading Pattern for Machine Learning
Performance Management to accelerate Cascading application development and management

- Plugin collects internal run-time and execution metadata directly from any Cascading application and visualizes data application
- Reduced cluster utilization costs
- Improved performance
- Immediate application failure diagnosis
Druid

Scalable and cost-effective time-series database

- Sub-second queries
- Ideal for user-facing analytics
- Can handle thousands of concurrent users
- Real-time streams
- Lock-free ingestion
- Query 10,000+ events per second per node
Questions?

Connect:
Email: manny@rebelai.com
Twitter: @epuentes | @Rebel_AI
Website: MannyPuentes.com | Rebel.AI
LinkedIn: http://www.linkedin.com/in/MannyPuentes