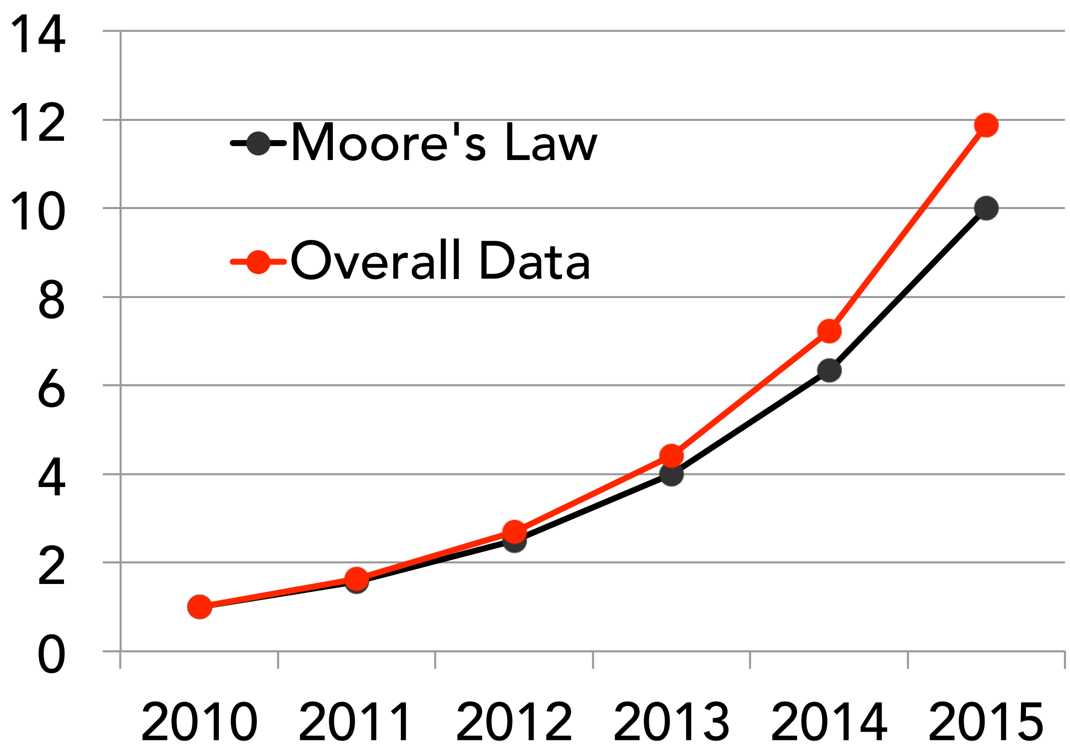
Turning Data into Value

Ion Stoica
CEO, Databricks
(also, UC Berkeley and Conviva)

Data is Everywhere

Easier and cheaper than ever to collect

Data grows faster than Moore's law



(IDC report*)

The New Gold Rush

Everyone wants to extract value from data

- » Big companies & startups alike

Huge potential

- » Already demonstrated by Google, Facebook, …

But, untapped by most companies

- » “We have lots of data but no one is looking at it!”

Extracting Value from Data Hard

Data is massive, unstructured, and dirty

Question are complex

Processing, analysis tools still in their “infancy”

Need tools that are

- » Faster
- » More sophisticated
- » Easier to use

Turning Data into Value

Insights, diagnosis, e.g.,

» Why is user engagement dropping?

» Why is the system slow?

» Detect spam, DDoS attacks

Decisions, e.g.,

» Decide what feature to add to a product

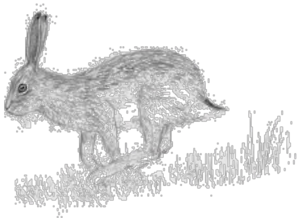
» Personalized medical treatment

» Decide when to change an aircraft engine part

» Decide what ads to show

Data only as useful as the decisions it enables

What do We Need?

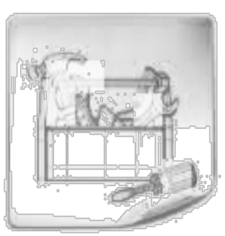


Interactive queries: enable faster decisions

» E.g., identify why a site is slow and fix it

Queries on streaming data: enable decisions on real-time data

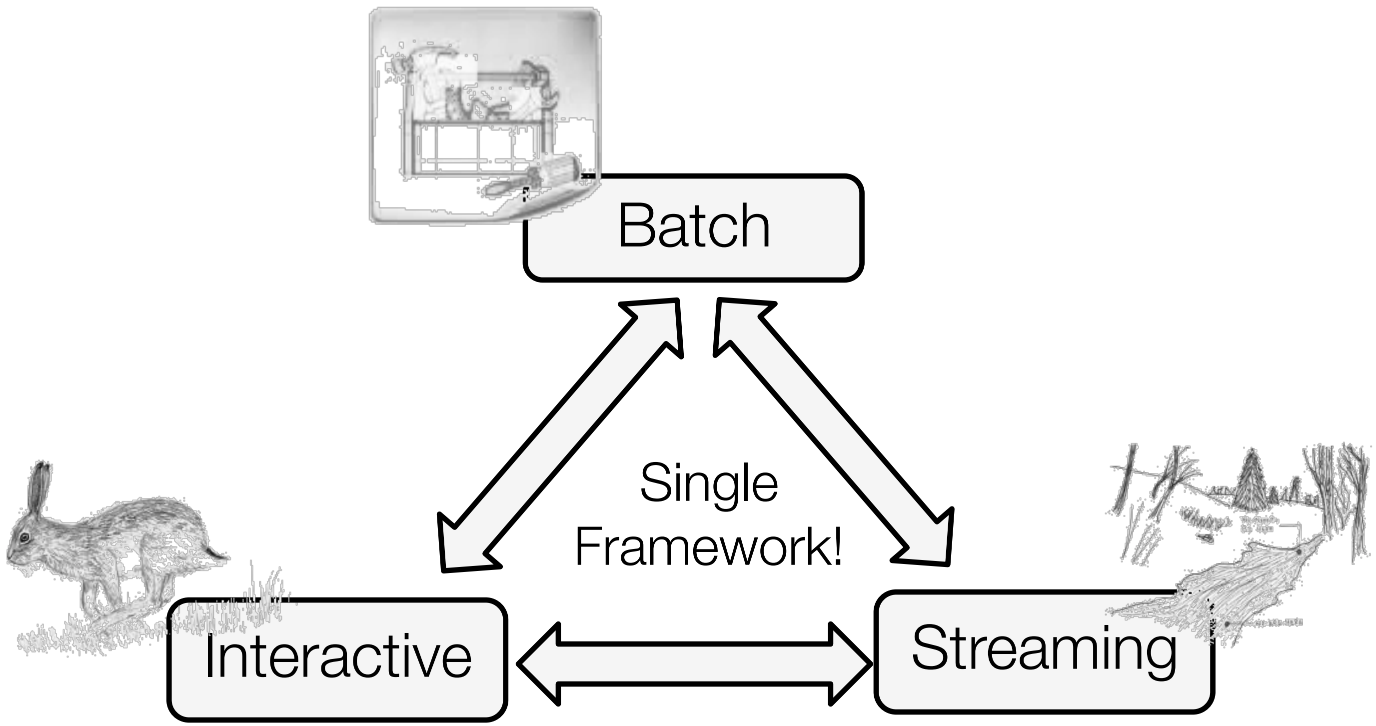
» E.g., fraud detection, detect DDoS attacks



Sophisticated data processing: enable "better" decisions

» E.g., anomaly detection, trend analysis

Our Goal



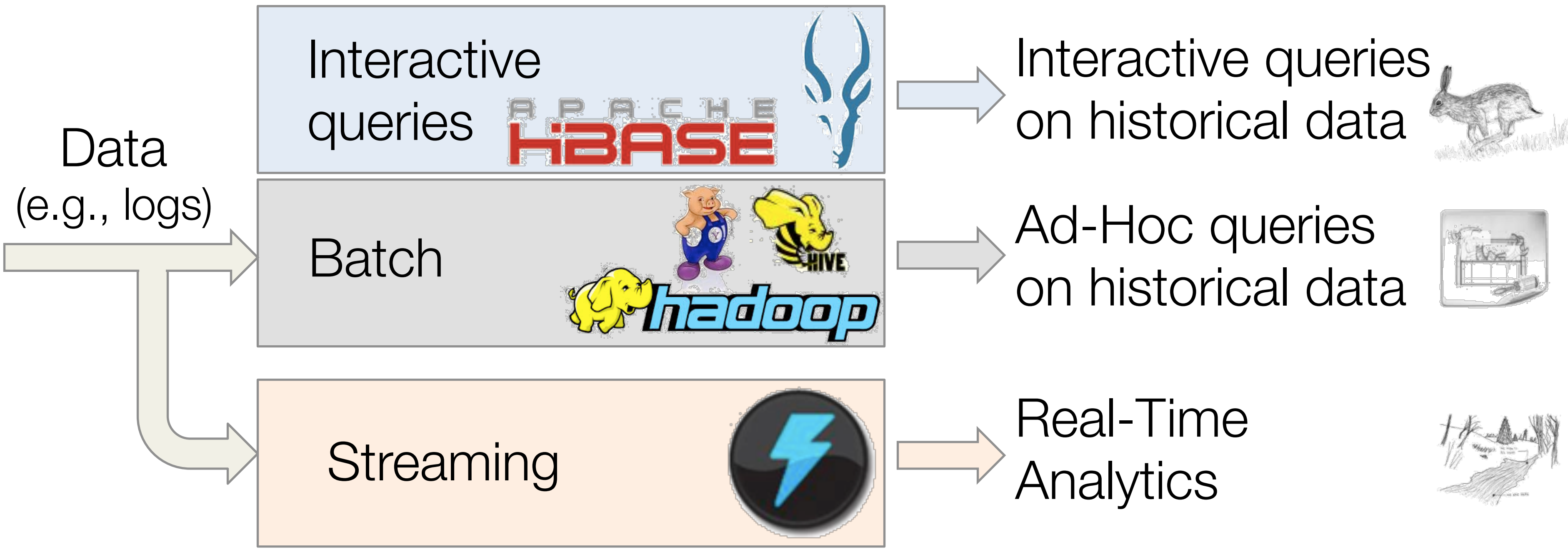
Support *batch*, *streaming*, and *interactive* computations…

… in a unified framework

Easy to develop *sophisticated* algorithms (e.g., graph, ML algos)

The Need For Unification

Today's state-of-art analytics stack

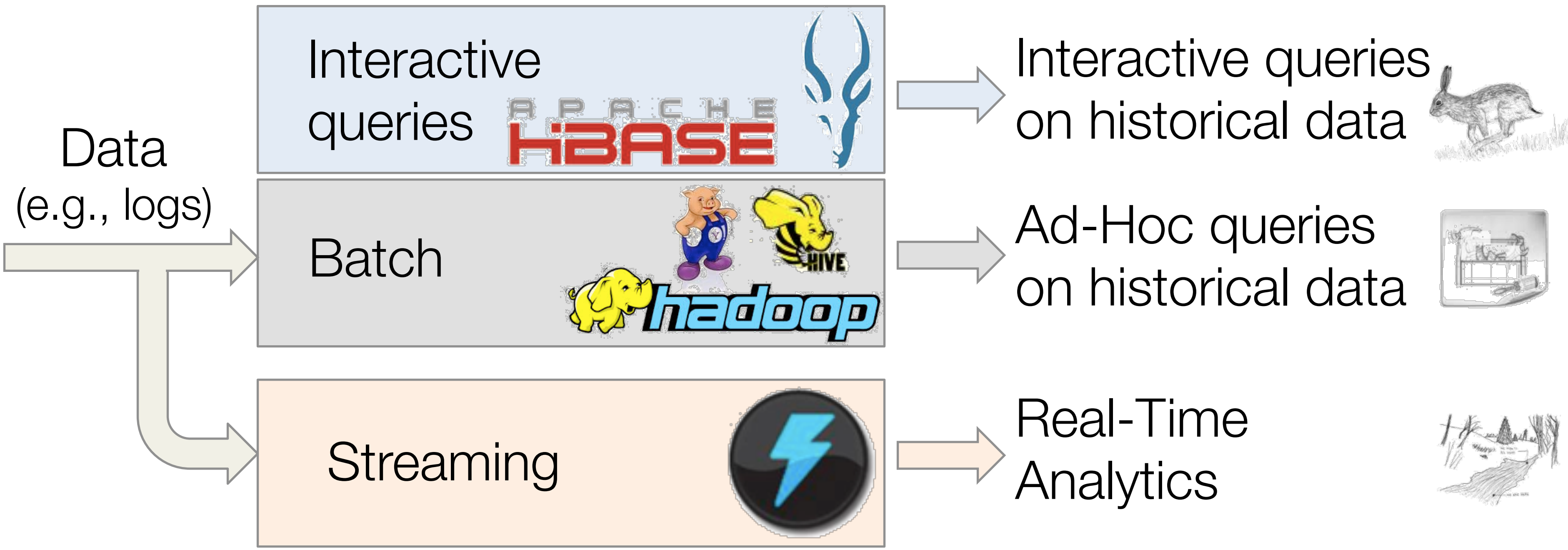


Challenge 1: need to maintain three stacks

- Expensive and complex
- Hard to compute consistent metrics across stacks

The Need For Unification

Today’s state-of-art analytics stack



Challenge 2: hard/slow to share data, e.g.,

» Hard to perform interactive queries on streamed data

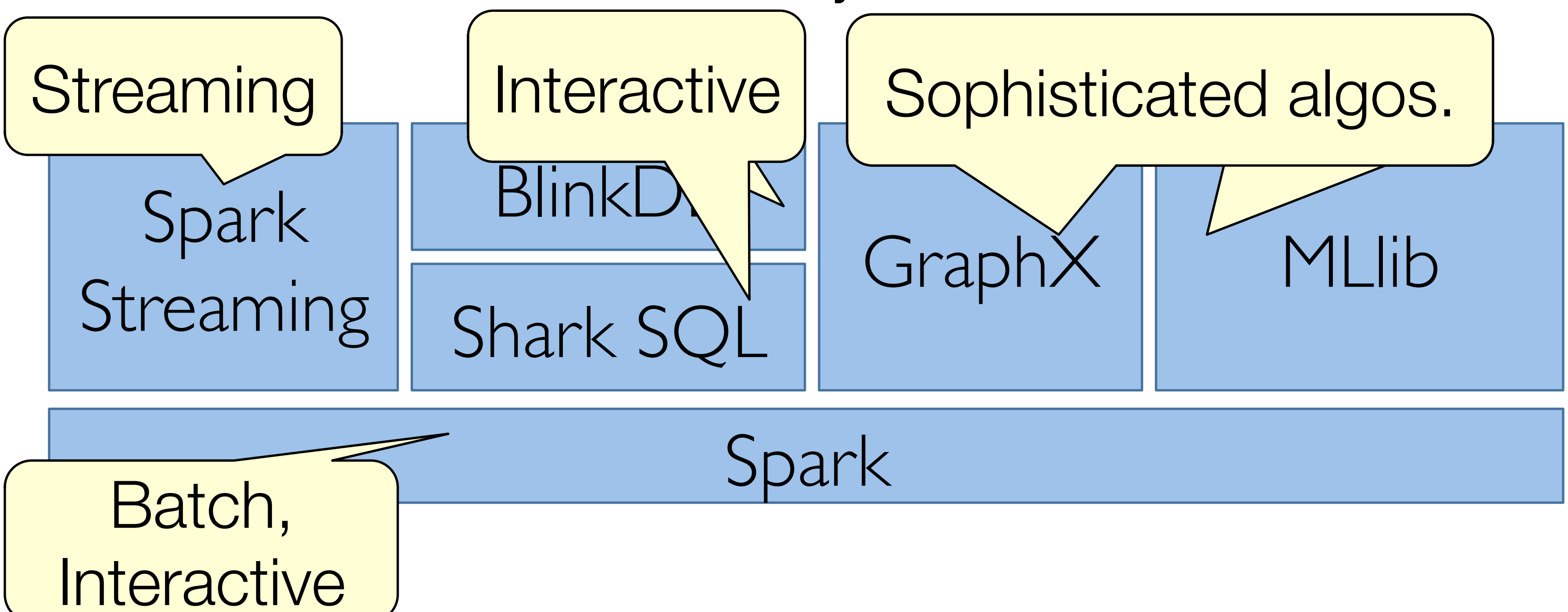
Spark

Unifies ***batch, streaming, interactive*** comp.

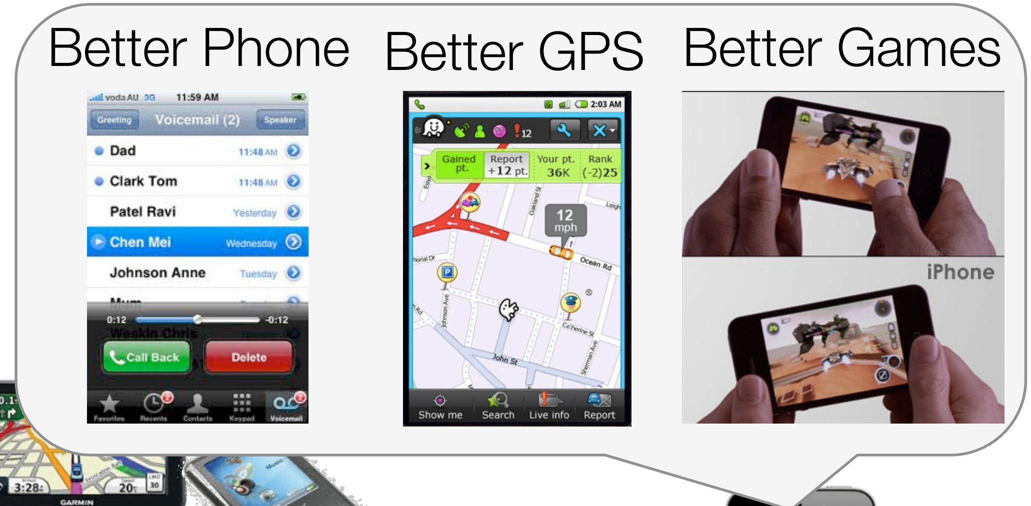
Easy to build sophisticated applications

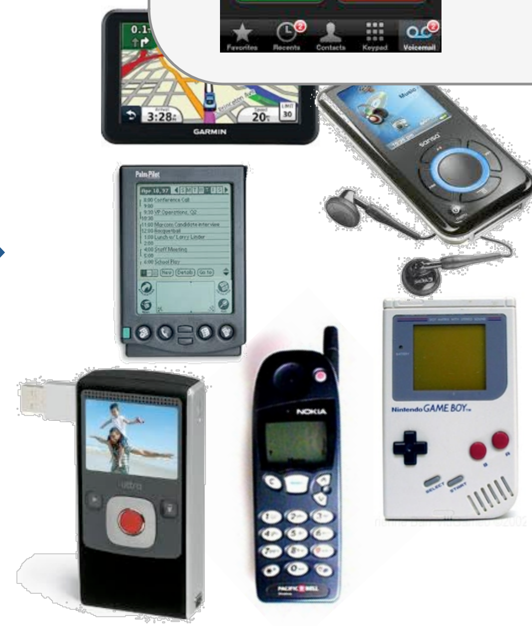
» Support iterative, graph-parallel algorithms

» Powerful APIs in Scala, Python, Java



An Analogy





First cellular phones

Specialized devices

Unified device (smartphone)

An Analogy

Batch processing

Specialized systems

Unified system

Turning Data into Value, Examples

Unify real-time and historical data analysis

- Easier to build and maintain
- Cheaper to operate
- Easier to get insights, faster decisions

Unify streaming and machine-learning

- Faster diagnosis, decisions (e.g., better ad targeting)

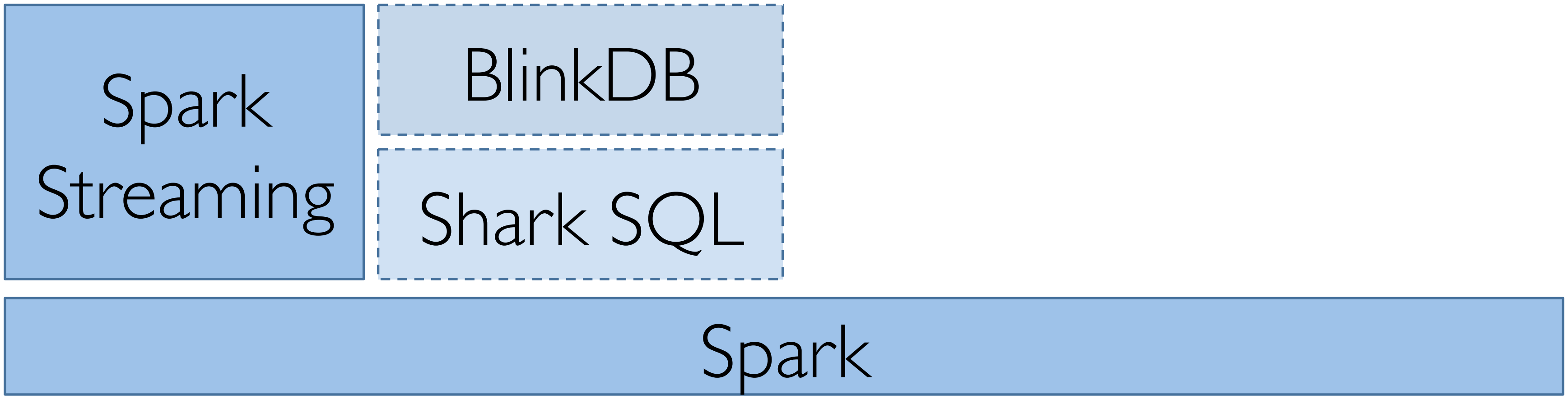
Unify graph processing and ETLs

- Faster to get social network insights (e.g., improve user experience)

Unify Real-time & Historical Analysis

Single implementation (stack) providing

- » Streaming
- » Batch (pre-computing results)
- » Interactive computations/queries



Unify Real-time & Historical Analysis

Batch and streaming codes virtually the same

» Easy to develop and maintain consistency

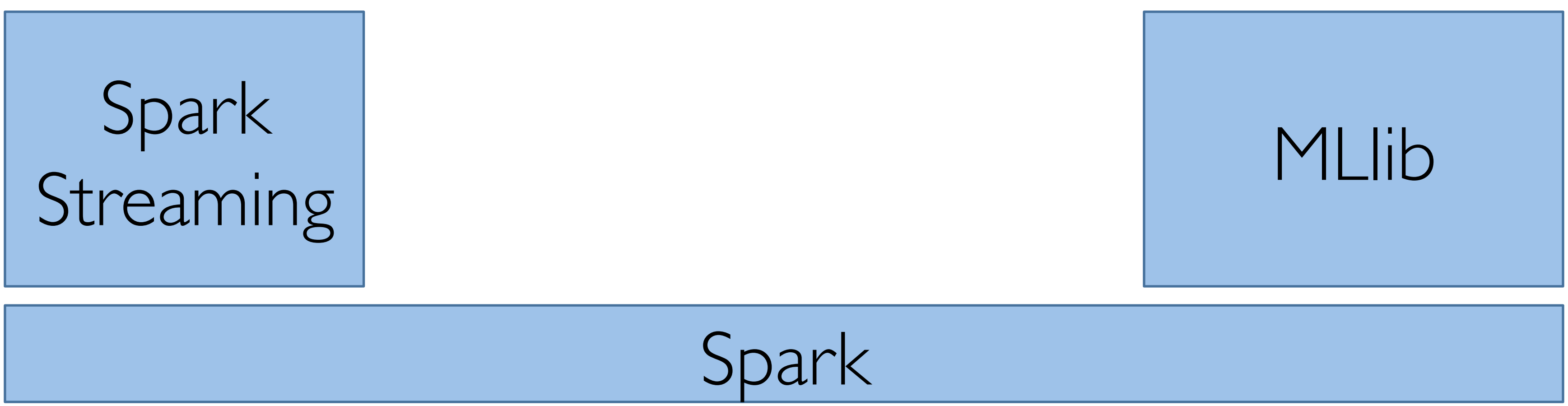
```
// count words from a file (batch)
val file = sc.textFile("hdfs://.../pagecounts-*.gz")
val words = file.flatMap(line => line.split(" "))
val wordCounts = words.map(x => (x, 1)).reduceByKey(_ + _)
wordCounts.print()
```

```
// count words from a network stream, every 10s (streaming)
val ssc = new StreamingContext(args(0), "NetCount", Seconds(10), ..)
val lines = ssc.socketTextStream("localhost", 3456)
val words = lines.flatMap(_.split(" "))
val wordCounts = words.map(x => (x, 1)).reduceByKey(_ + _)
wordCounts.print()
ssc.start()
```

Unify Streaming and ML

Sophisticated, real-time diagnosis & decisions, e.g.,

- » Fraud detection
- » Detect denial of service attacks
- » Early notification of service degradation and failures

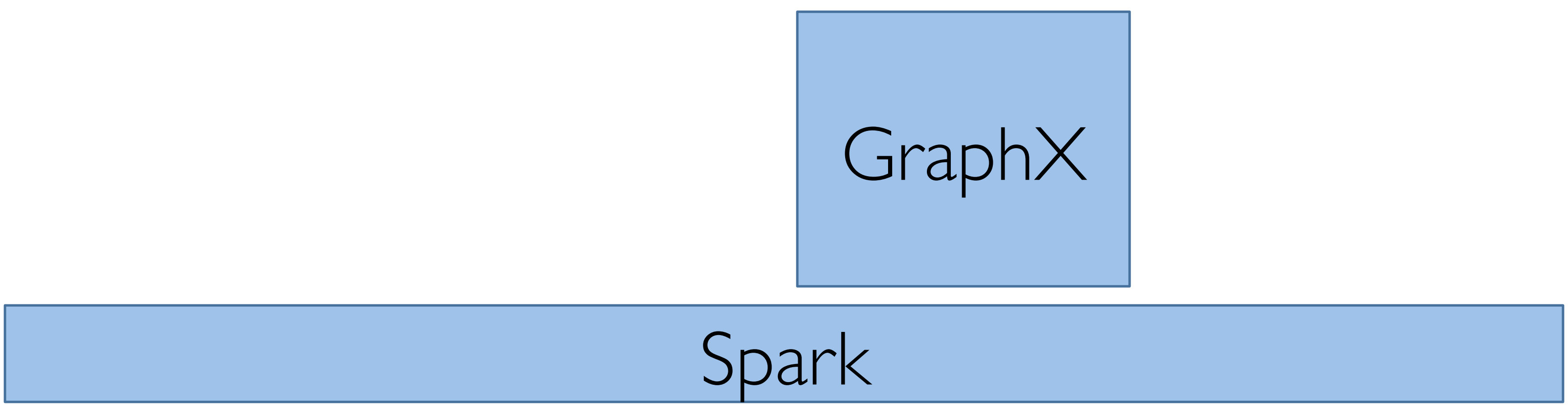


Unify Graph Processing and ETL

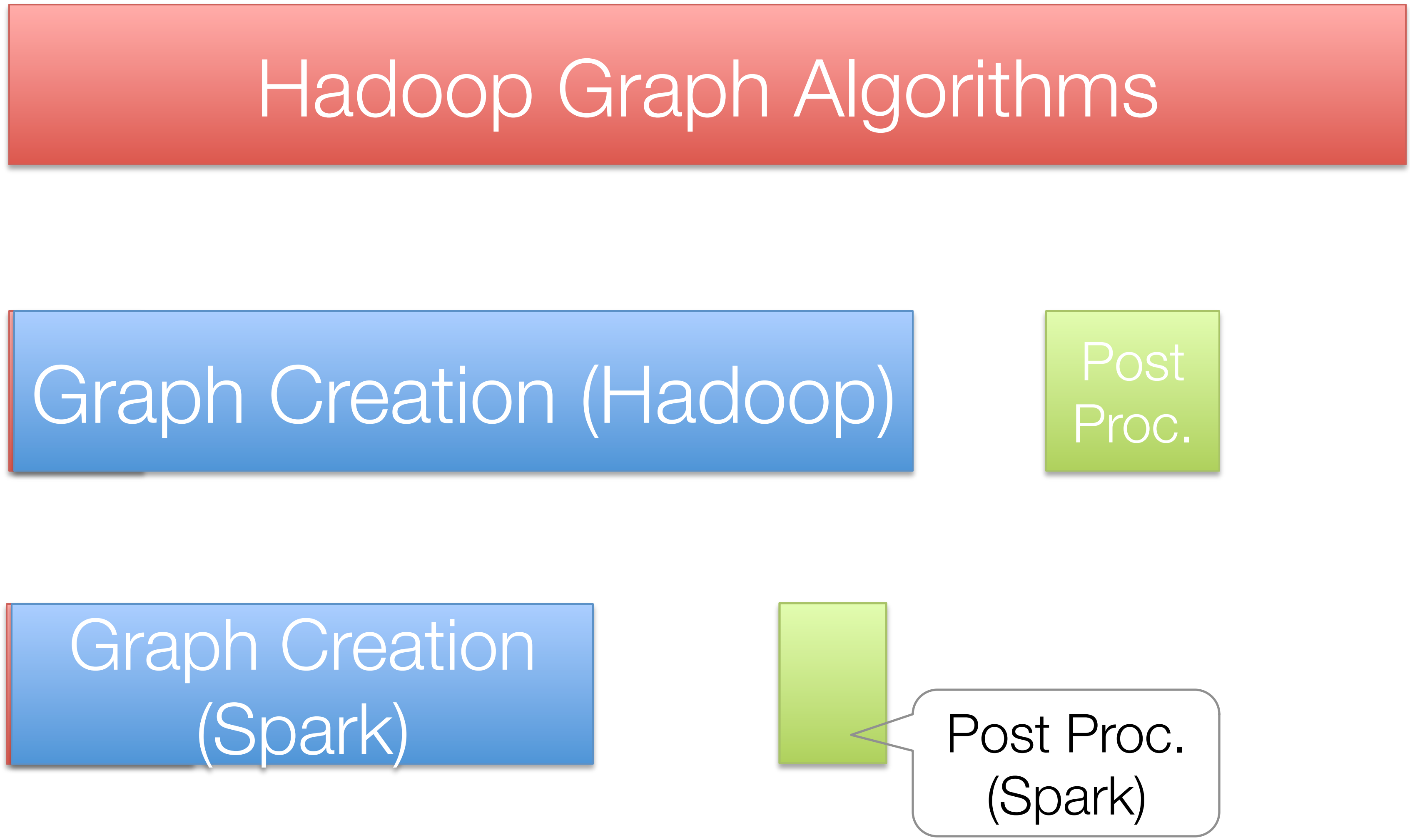
Graph-parallel systems (e.g., Pregel, GraphLab)

- » Fast and scalable, but…
- » … inefficient for graph creation, post-processing

GraphX: unifies graph processing and ETL



Unify Graph Processing and ETL



“Crossing the Chasm”

amazon web services

AWS Products & Solutions

Databricks aims to build next-generation analytic tools for Big Data

A new startup will accelerate the maturation of the Berkeley Data Analytics Stack

ca | @bigdata | Comment | September 25, 2013

New Cloudera Partner Program Harnesses Power of Innovative Startups Databricks, the Inaugural Partner of Cloudera Connect: Innovators, Teams With Cloudera for High-Speed Data Analytics

© Marketwire 2013

2013-10-28 12:10:03 -

Print article

WANdisco Announces Support for In-Memory Data Processing Technologies, Spark and Shark

MARKET WIRED

Press Release: WANdisco, Plc. – Wed, Jun 26, 2013 9:00 AM EDT

Cloudera Partnership

Integrate Spark with Cloudera Manager

Spark will become part of CDH

Enterprise class support and professional services available for Spark

We are Committed to…

… open source

» We believe that any successful analytics stack will be open source

… improve integration with Hadoop

» Enable every Hadoop user take advantage of Spark

… work with partners to make Apache Spark successful for enterprise customers

databricks™

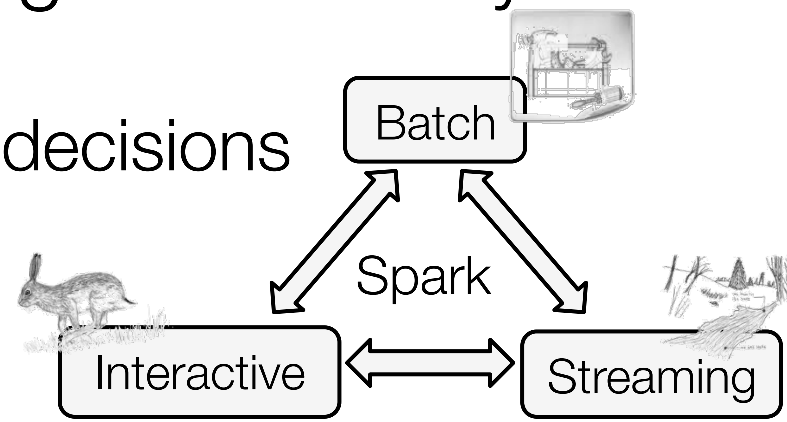
Summary

Everyone collects but few extract value from data

Unification of comp. and prog. models key to

» Efficiently analyze data

» Make sophisticated, real-time decisions



Spark is unique in unifying

» batch, interactive, streaming computation models

» data-parallel and graph-parallel prog. models

Many use cases in the rest of the program!