Welcome to the first Workshop on Big data Open Source Systems (BOSS)

September 4th, 2015

Co-located with VLDB 2015

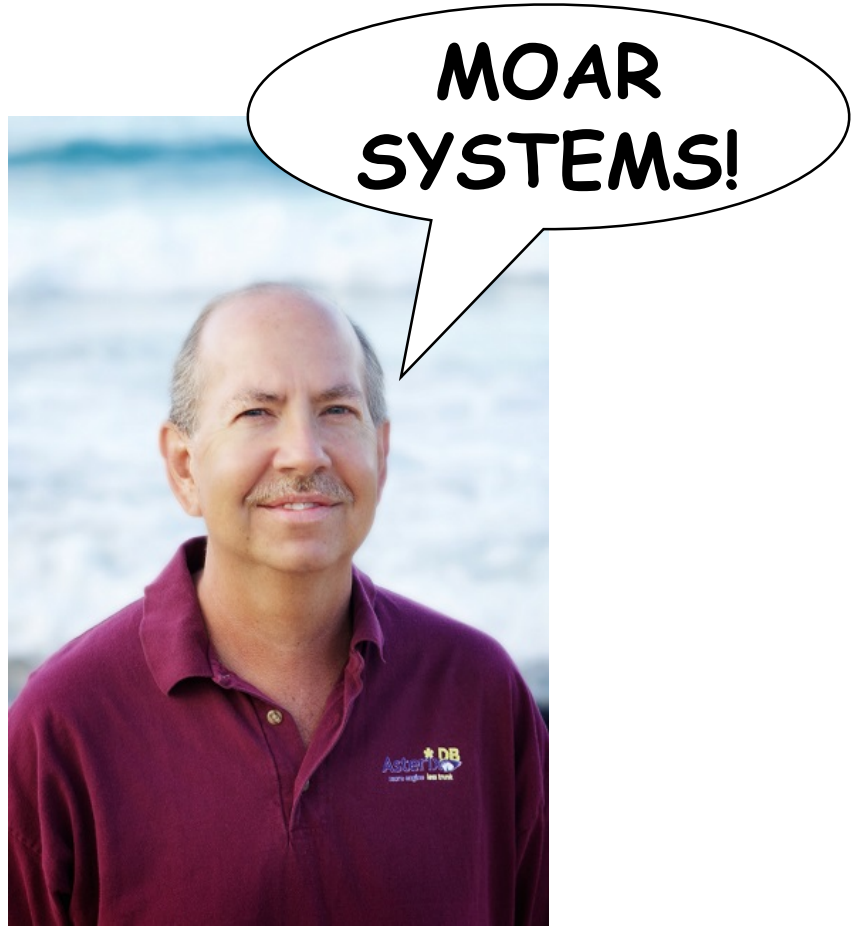
Tilmann Rabl

Hands on Big Data

- 8 parallel tutorials
- 8 systems
  - Open source
  - Publicly available
- Presenters
  - System experts
- Hands on
  - This is not a demo!
- You can pick two!

But why?

- Initial idea: Malu Castellanos
- Mike Carey
  - Doing It On Big Data: a Tutorial/Workshop
  - Driving force
- Other people involved
  - Volker Markl
  - Norman Patton
  - Lipyeow Lim
  - Kerstin Forster
- Experiment
  - Tell us what you think
  - Email: rabl@tu-berlin.de



Presented Systems

- Apache AsterixDB

- Apache Flink

- Apache Reef

- Apache Singa

- Apache Spark

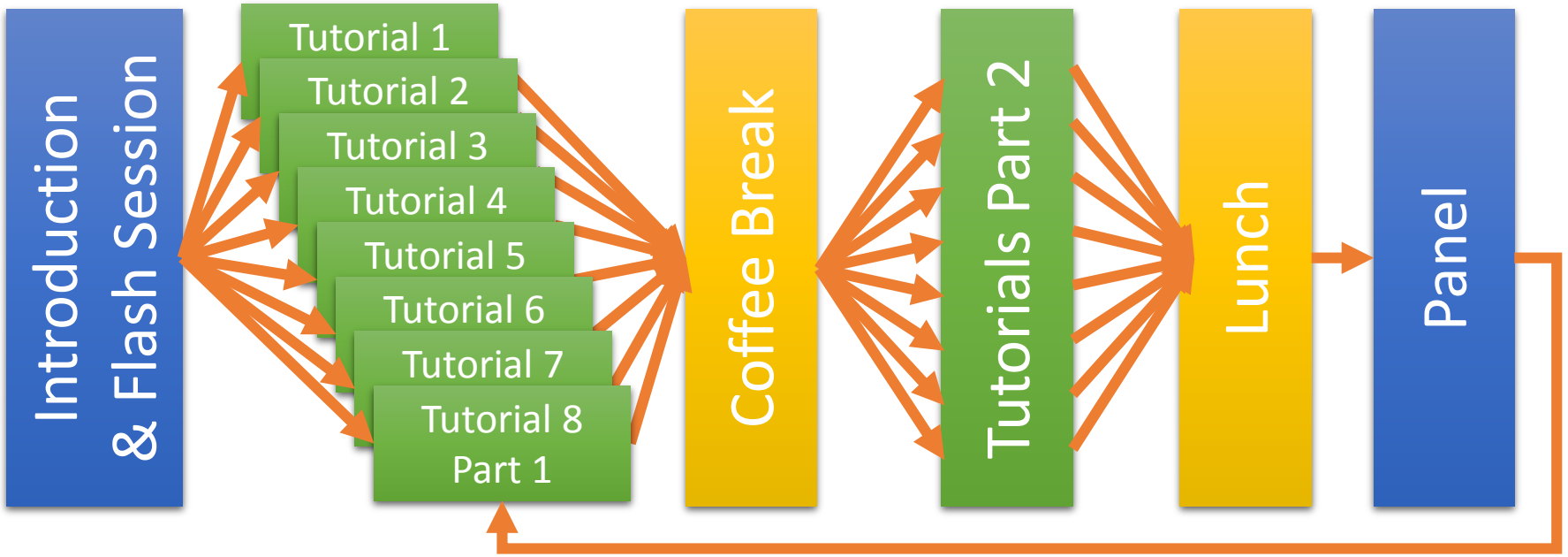
- Padres

- rasdaman

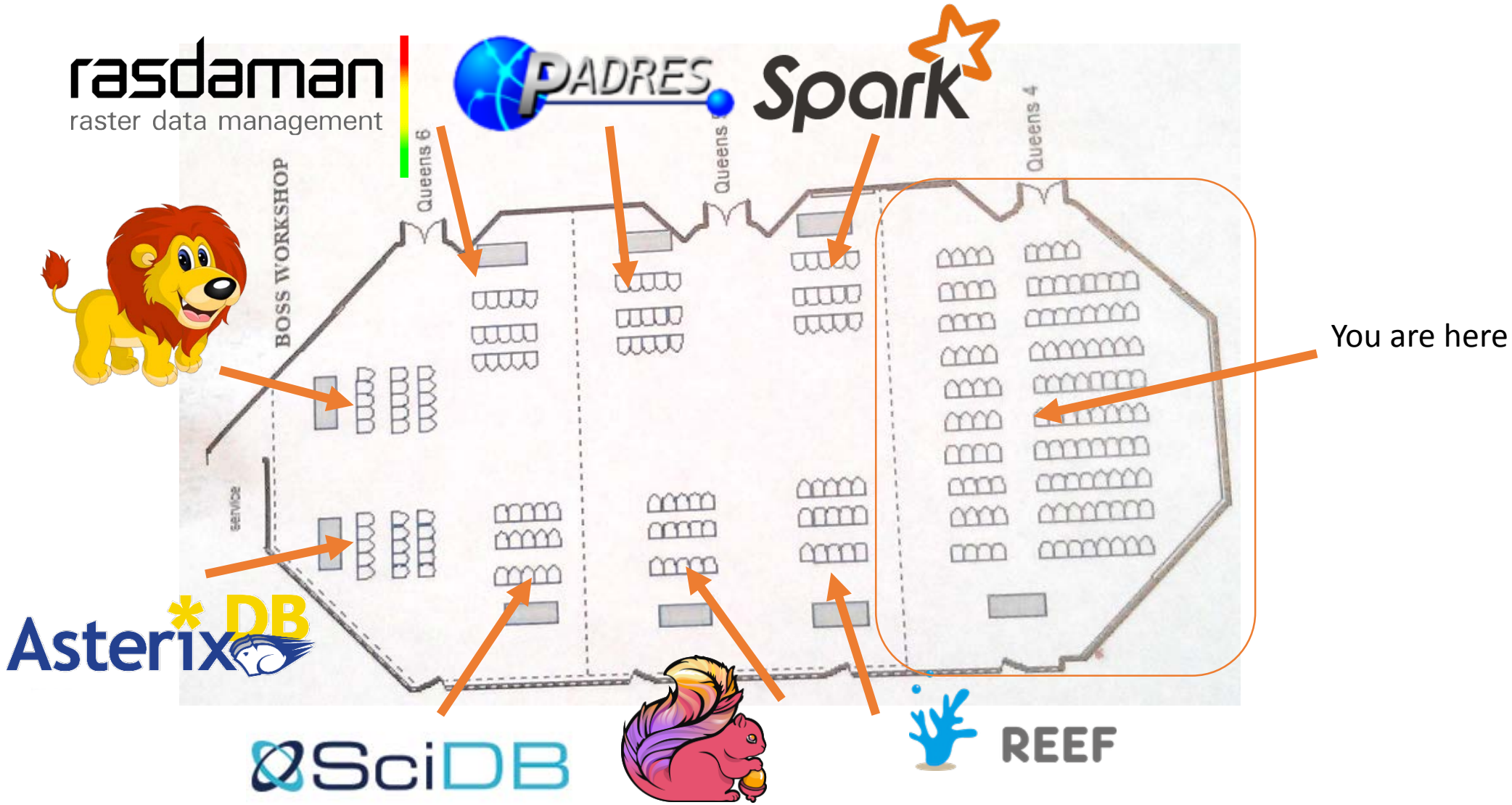
- SciDB

Massively Parallel Program

- Bulk Synchronous Parallel

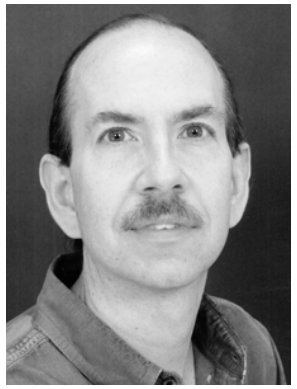


Runtime Environment

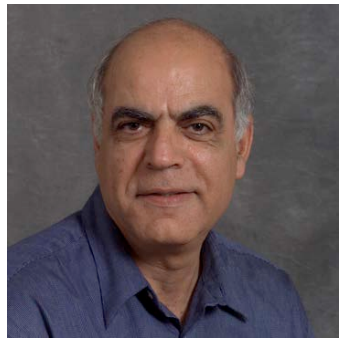


Panel – Big Data and Exascale

- Panel Chair
  - Chaitanya Baru, San Diego Supercomputing Center
- Panelists
  - Arie Shoshani, LBNL
  - Guy Lohmann, IBM
  - Mike Carey, UC Irvine
  - Paul G. Brown, Paradigm4
  - Peter Baumann, Jacobs University
  - Volker Markl, TU Berlin



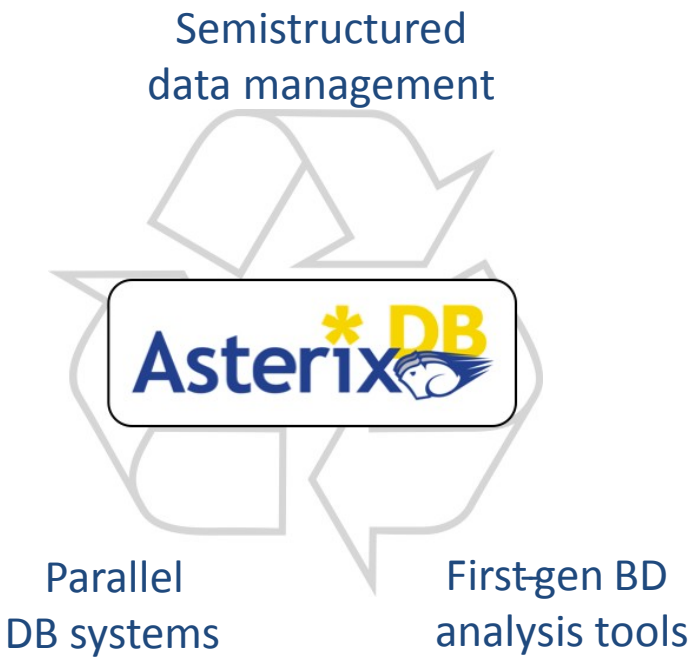



Apache AsterixDB (Incubating)

AsterixDB: “One Size Fits a Bunch!”

Wish-list:

- Able to manage data
- Flexible data model
- Full query capability
- Continuous data ingestion
- Efficient and robust parallel runtime
- Cost proportional to task at hand
- Support today’s “Big Data data types”



Apache Flink

Apache Flink™: Stream and Batch processing at Scale

Marton Balassi (ELTE/SZTAKI, Hungary)
Paris Carbone (KTH, Stockholm, Sweden)
Gyula For a (SICS, Stockholm, Sweden)
Vasia Kalavri (KTH, Stockholm, Sweden)
Asterios Katsifodimos (TU Berlin, Germany)

What is Flink?

Applications

Hive | Cascading | Giraph

Mahout | Pig | Crunch

Data processing engines

MapReduce  | Flink 

Spark  | Storm  | Tez 

App and resource management

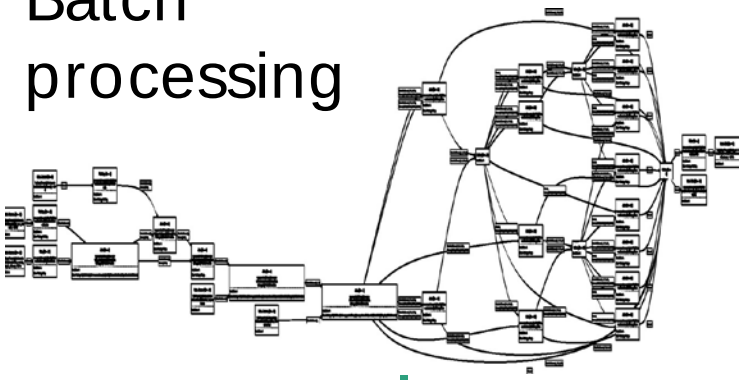
Yarn | Mesos

Storage, streams

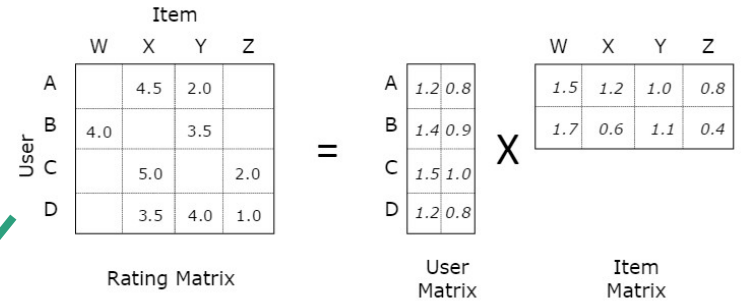
HDFS | HBase | Kafka | ...

What can I do with Flink?

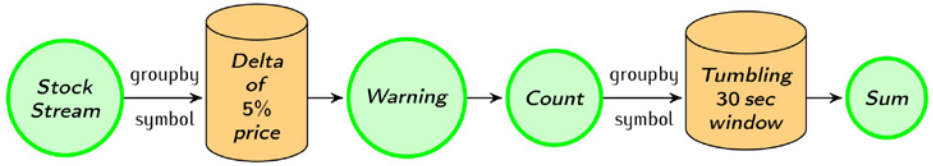
Batch processing



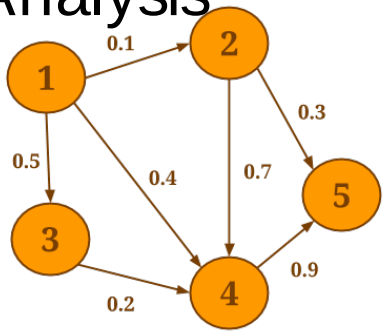
Machine Learning at scale



Stream processing



Graph Analysis

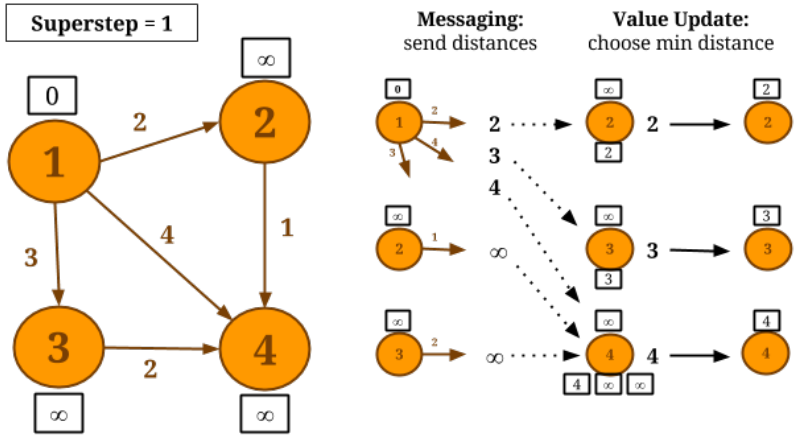


Flink

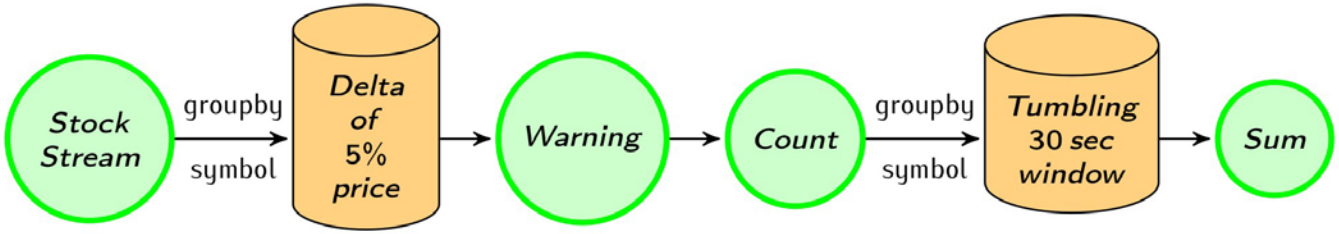
An engine that can **natively** support all these workloads.

But what will I do with Flink today?

- **Graph processing**
  - ETL on Datasets
  - Graph creation & analysis



- **Stream Processing**
  - Rolling Aggregates
  - Windows & Alerts



Agenda

- **Introduction**
  - 15’ Overview
  - 15' Gelly (Graph) API
- **30' Break**
- **Graph Processing**
  - 20' DataSet/Gelly Hands-on
- **Stream processing with Flink**
  - 10’ DataStream API
  - 15’ Fault Tolerance Demo
  - 45' Streaming Hands-on

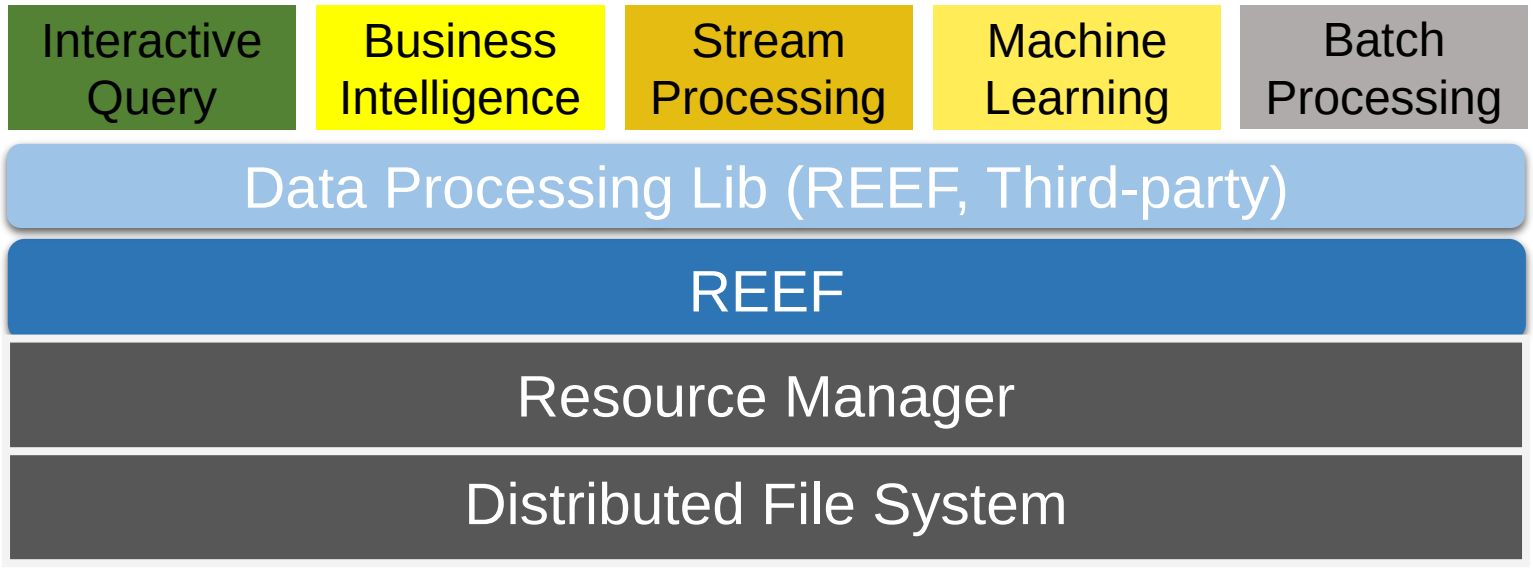
Apache Reef

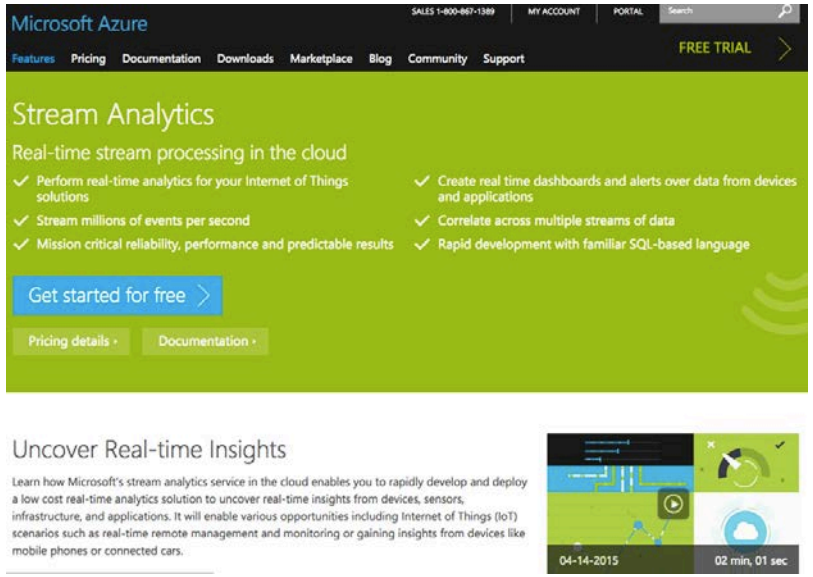
Deep Dive into Apache REEF (Incubating)

Byung-Gon Chun, Brian Cho (Seoul National University)

BOSS 2015
Sep. 4, 2015

A meta-framework that eases the development of Big Data applications atop resource managers such as YARN and Mesos





- ✓ Reusable control plane for coordinating data plane tasks
- ✓ Adaptation layer for resource managers
- ✓ Container and state reuse across tasks from heterogeneous frameworks
- ✓ Simple and safe configuration management
- ✓ Scalable local, remote event handling
- ✓ Java and C# (.NET) support

In production use (Microsoft Azure)

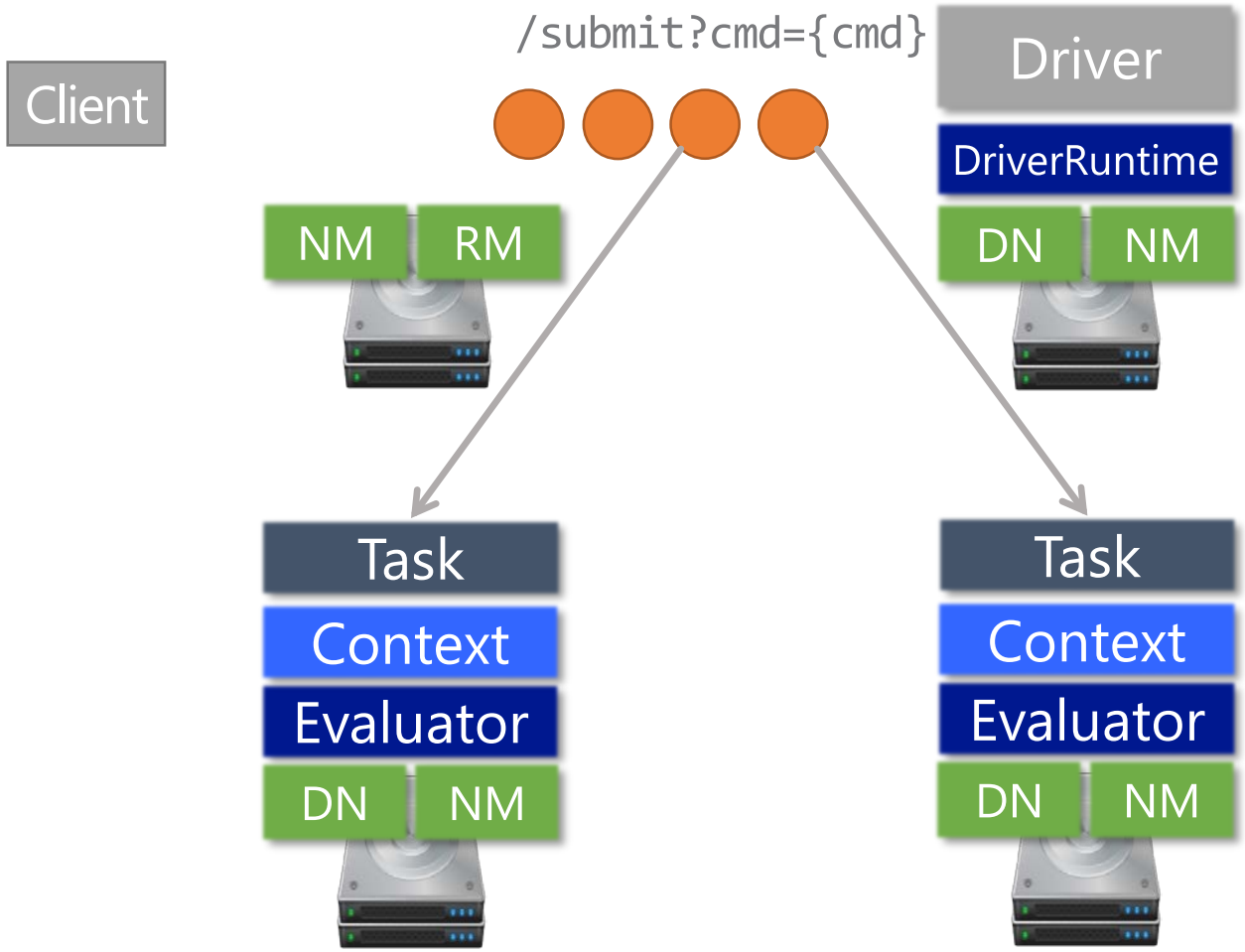
Deep Dive into Apache REEF (Incubating)

Byung-Gon Chun, Brian Cho (Seoul National University)

BOSS 2015
Sep. 4, 2015

Tutorial

1. What is REEF?
2. Install REEF
3. Run your first REEF job: HelloREEF
4. Why would you want REEF?
5. Create your own Task Scheduler with REEF



Contact: Byung-Gon Chun bgchun@gmail.com
Brian Cho chobrian@gmail.com

Apache Singa

Apache SINGA

Apache incubator ™

A General Distributed Deep Learning Platform

- Motivation
  - Deep learning is effective for classification tasks, e.g., image recognition
  - Training code is complex to write from the scratch
  - Training is time consuming, e.g., 10 days or weeks
- Goals
  - Easy to use
    - General to support popular deep learning models
    - Extensible for users to do customization, e.g., training new models
  - Scalable
    - Reduce training time with more computation resources, e.g. machines
    - Improve efficiency of one training iteration by synchronous training
    - Reduce total number of training iterations by asynchronous training

Apache Spark

Spark Tutorial

Reynold Xin @rxin

Sep 4, 2015 @ VLDB BOSS 2015

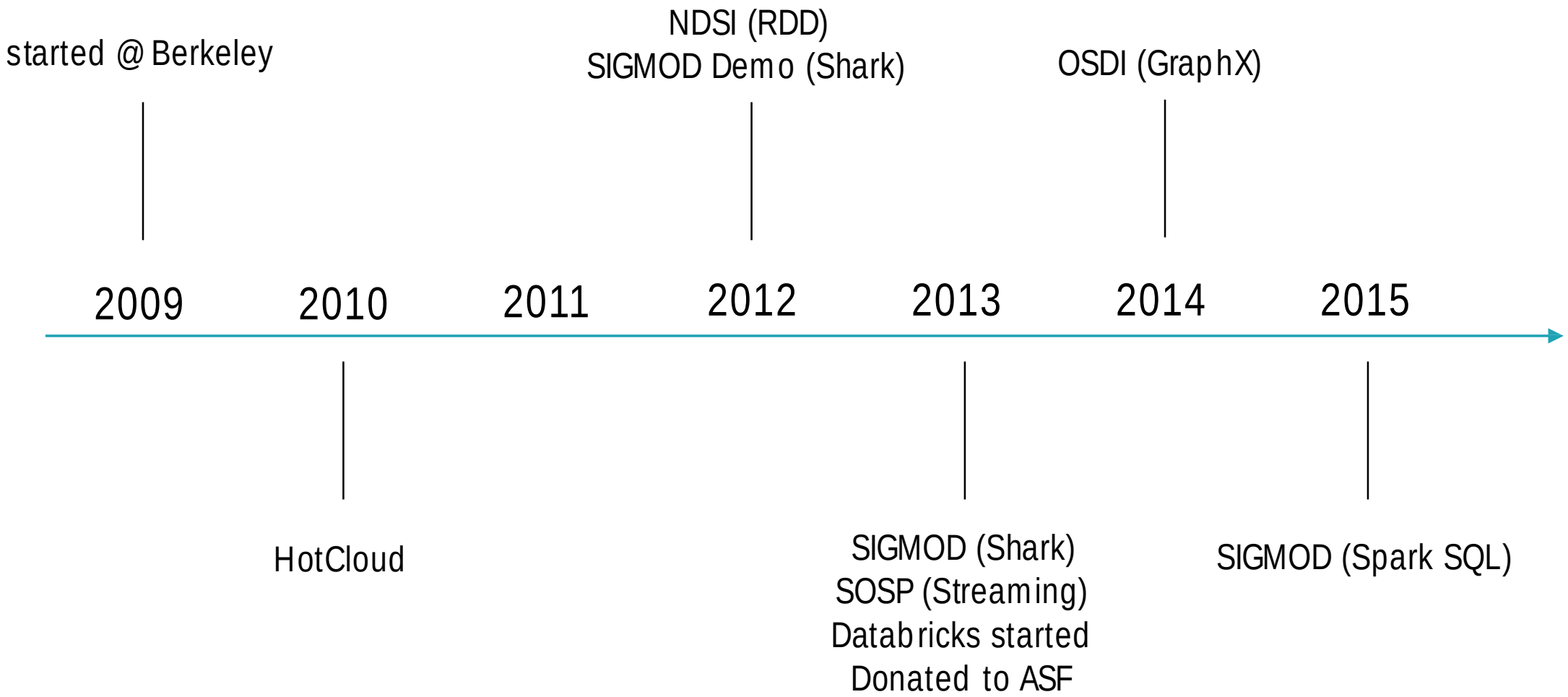
Apache Spark

Fast & general distributed data processing engine, with APIs in SQL, Scala, Java, Python, and R

800+ contributors and many academic papers

Largest open source project in (big) data & at Apache

A Brief History



Users

1000+ companies

...

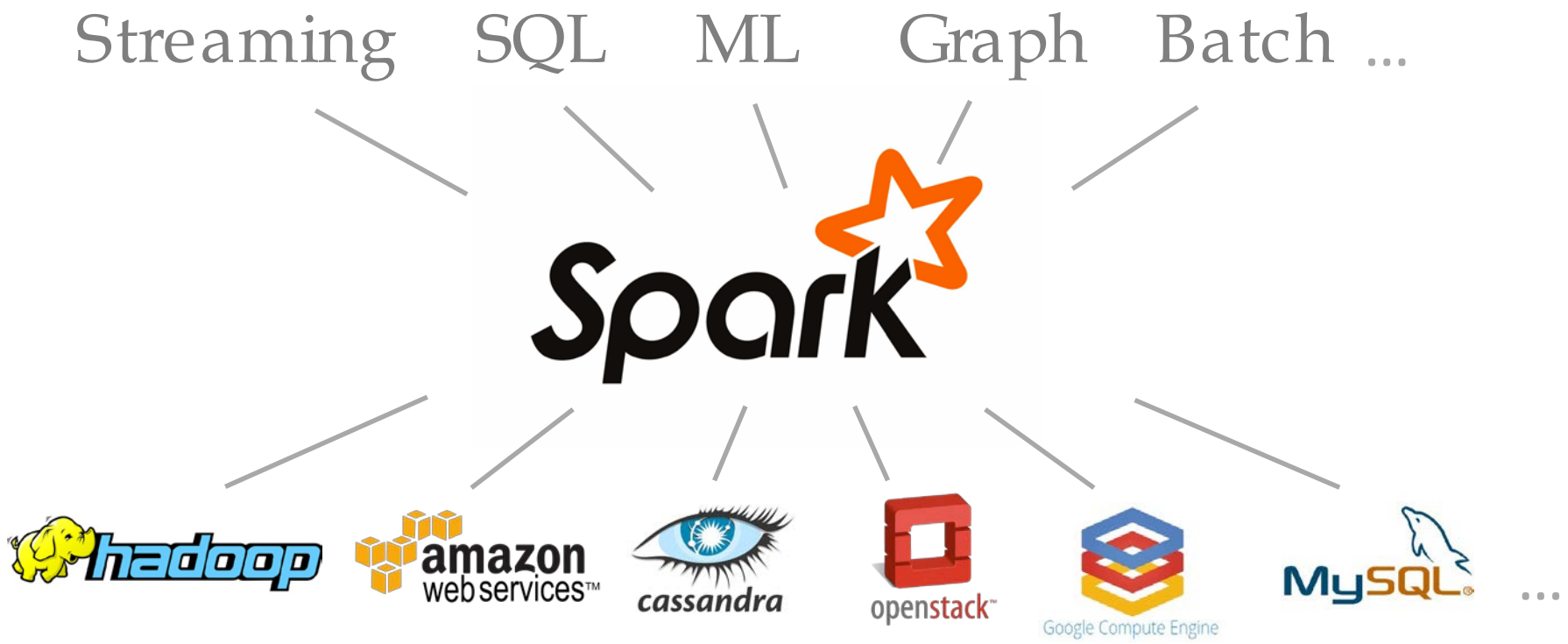
Distributors + Apps

50+ companies

...

Our Goal for Spark

Unified engine across data workloads and platforms



Agenda Today

Spark 101: RDD Fundamentals

Spark 102: DataFrames

Spark 201: Understanding Spark Internals

(with exercises in Databricks notebooks)

PADRES

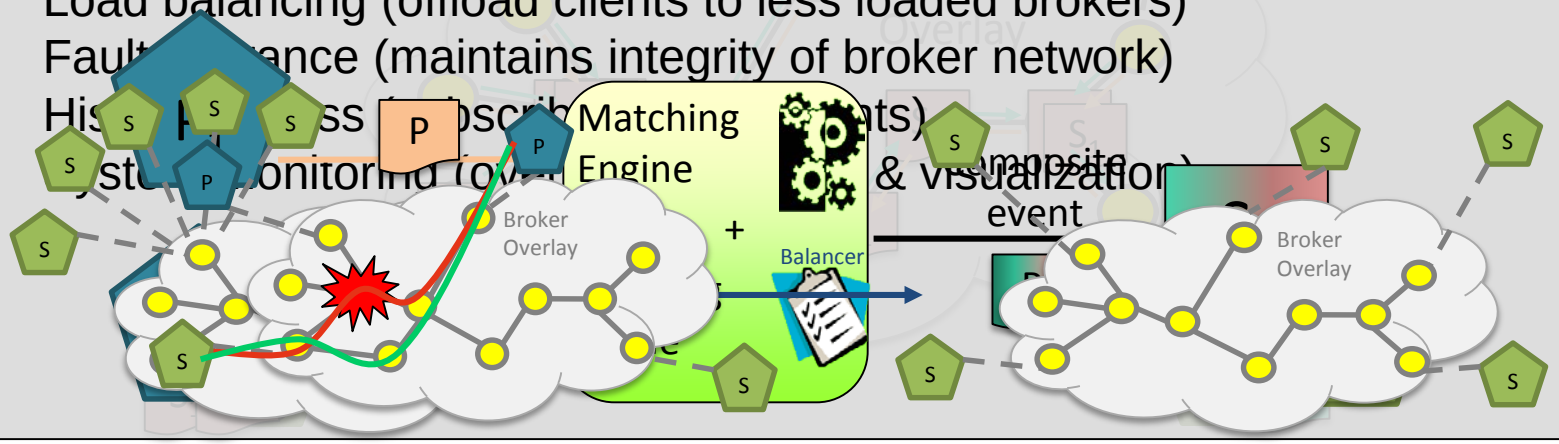
Presenter: Kaiwen Zhang
University of Toronto

Pub/Sub is a communication paradigm / middleware

Communication between information producers (publisher) and consumers (subscriber) is mediated by a set of brokers (p2p overlay).

Features

- Content-based routing
- Composite subscription (event P_1 and event P_2 occurred within 2s)
- Load balancing (offload clients to less loaded brokers)
- Fault [illegible]ance (maintains integrity of broker network)
- His[illegible]ss [illegible]ts)
- [illegible]onitorin[illegible] & visualization)



rasdaman

rasdaman

the Array Database

the pioneer Array DBMS: analytics on n-D dense/sparse arrays
optimization & parallel QP on multicore, cloud, modern hw
scalable from cubesat to datacenter federations
seamless integration with R, python, ...
operationally deployed on Petascale, basis for ISO Array SQL

www.rasdaman.org

SciDB

SciDB:

No cute animals …

No 5 color marketing brochure …

… just an …

Open Source,

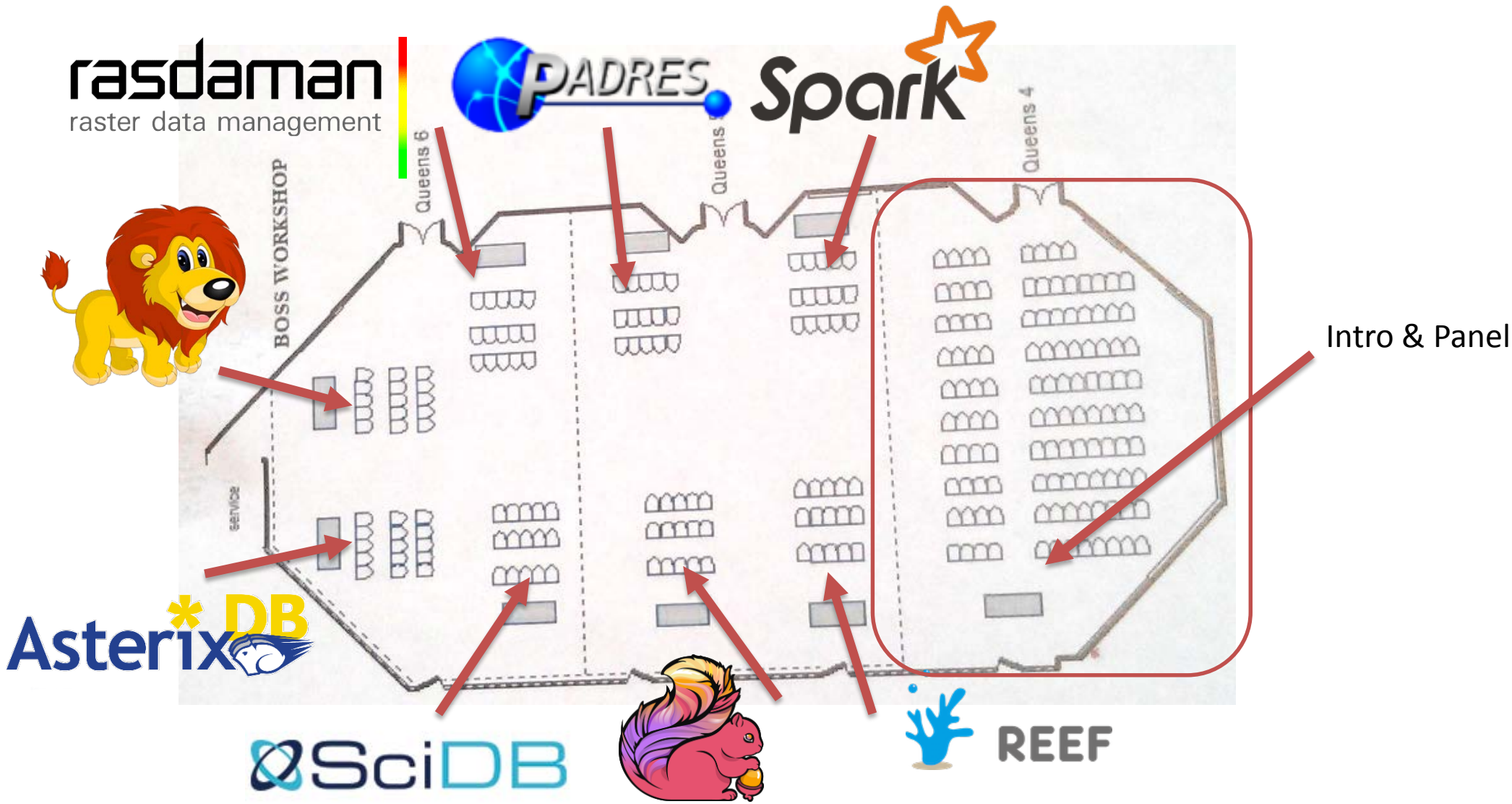
Transactional,

Massively Parallel,

Array DBMS with

A Scalable Analytic Query Engine.

Let‘s go!


rasdaman
raster data management
PADRES
Spark
BOSS WORKSHOP
Queens 6
Queens 4
service
AsterixDB
SciDB
REEF
Intro & Panel