Internal Data Needs

- Process 100s of Billions events a day
- Wide variety of uses
  - Targeting, Personalization, Abuse prevention, BI and Analytics, etc.
  - Support 1000s of internal users
  - Support 100s of internal products
- Focus on
  - Match the user’s needs with the right data set
  - Speed
  - Completeness
  - Accuracy
What’s the data menu?

▪ My data, my way!
  › M/R, Hive, Druid, Kafka, Spark, Storm, etc.

▪ Access controls according to needs and skills
  › Novice internal data users gets access to dashboards
  › Advanced internal users get access to more data

▪ Data governance
  › Good documentation
  › Handle access control, security, privacy, compliance

Data Ecosystem needs to cater to all of these access methods
Data Systems Overview

- **Proprietary Instrumentation SDKs**
- **Data Collection**

**Hadoop Grids**
(M/R, Hive, Pig, Spark)
- **Batch Processing Systems** → **Grid Feeds**
- **Real Time Grids**
(Storm, Spark)
- **Stream Processing Systems**
- **Batch Views**
- **Real time Views**
- **Kafka**

**Bi Tools** (OLAP)
- ML, Data Mining
- Webapps
- Data APIs
- Streaming Applications
Handle the lambda fork

- Stream and Batch
  - A minefield of code duplication and redundant work
- Approaches
  - Accept Duplication
  - Dynamic “DSL”
  - Push everything to the stream
Handle the lambda fork

- Unifying the view
  - Different Stream and Batch systems!?  
- Merged view, e.g.
  - Hive on batch and “mini-batch” real-time
  - Druid unifies batch and realtime natively
  - Other systems can choose which to connect to
Things to think about - now!

- Scale
- No downtime deploys
  - Real-time systems == “available now”
- New sources of data, “wider” data
  - How to dynamically accommodate?
- Open source
  - Balance being a user and a contributor
Thank You
System Overview

Client layer

- BI Tools (OLAP)
- WebApps
- Data APIs

Processing layer

- Native Map/Reduce, PIG
- RT query, ML, data mining
- Adhoc query, dashboard
  - Query IFs (J/O)DBC
  - Hive
  - Stream Processing (Storm)
    - Druid
    - Shark
    - Spark

- YARN
- Hadoop HDFS + MapReduce

Data collection

- Data Movement / Collection
- Proprietary Instrumentation SDKs
Lambda Architecture

Proprietary Instrumentation SDKs
Data Collection

Hadoop Grids
Batch Processing Systems

Grid Feeds
Batch Views
Real time Views

Real Time Grids
Stream Processing Systems

Kafka
Kappa Architecture

- Proprietary Instrumentation SDKs
- Data Collection
- Stream Processing
- Real Time Grids
  - Real time Views
- Additional Batch Processing
- Hadoop Grids
  - Batch Views
  - Grid Feeds