Data @ Yahoo

Sundeep Narravula and Eric Tschetter

{sundeepn,cheddar}@yahoo-inc.com

May 19, 2015

Internal Data Needs

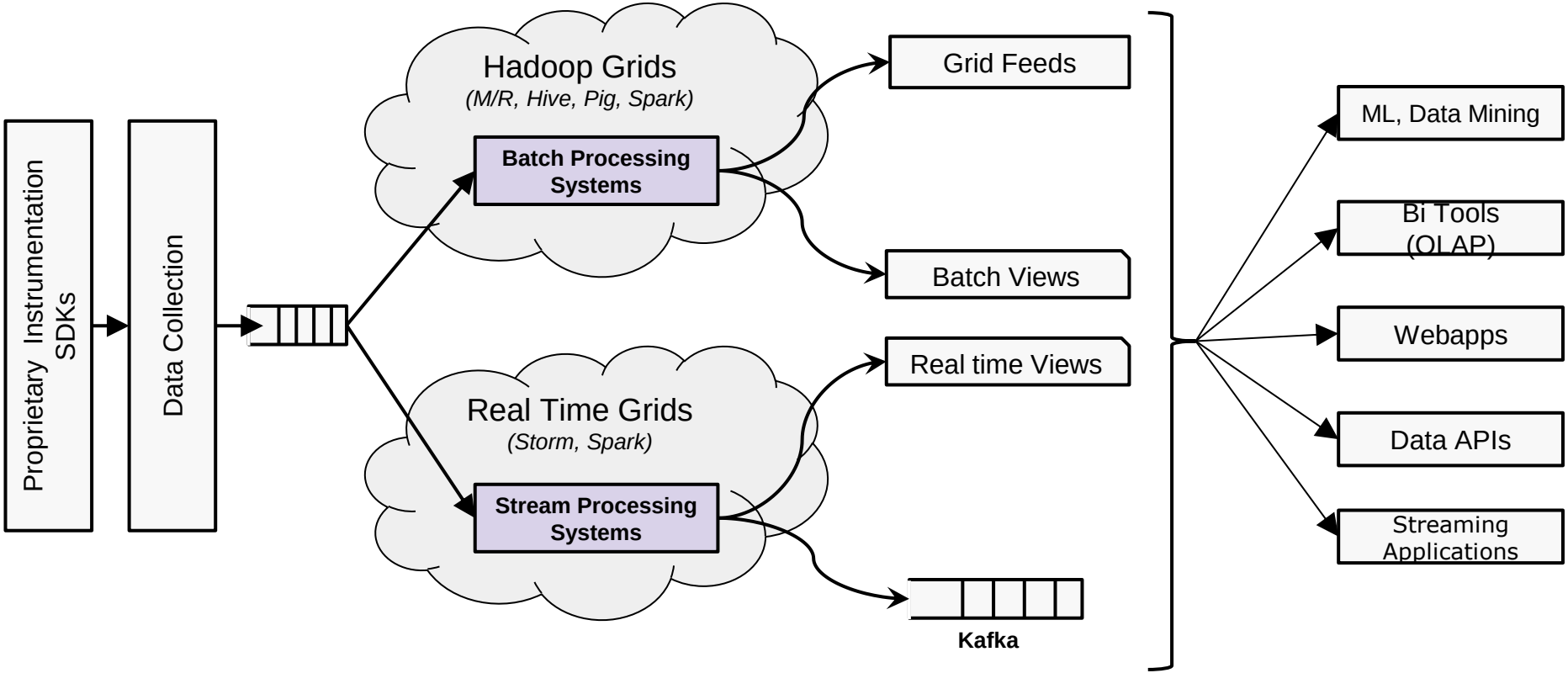
- Process 100s of Billions events a day
- Wide variety of uses
  - Targeting, Personalization, Abuse prevention, BI and Analytics, etc.
  - Support 1000s of internal users
  - Support 100s of internal products
- Focus on
  - Match the user's needs with the right data set
  - Speed
  - Completeness
 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



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

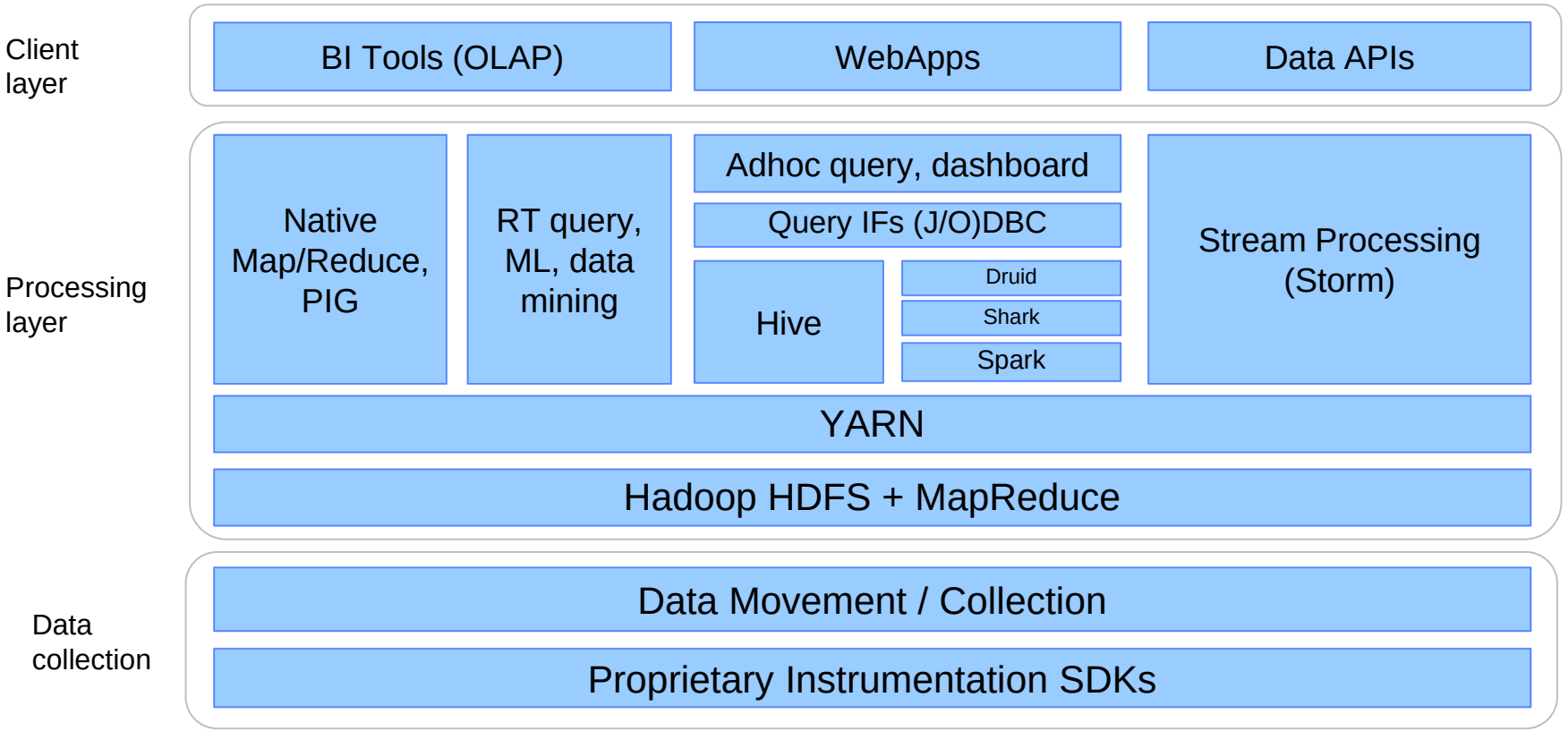
- Unifying the view
  - Different Stream and Batch systems!?
- Merged view, e.g.
  - Hive on batch and “mini-batch” real-time
  - Druid unifies batch and realtime natively
  - Other systems can choose which to connect to

Things to think about - now!

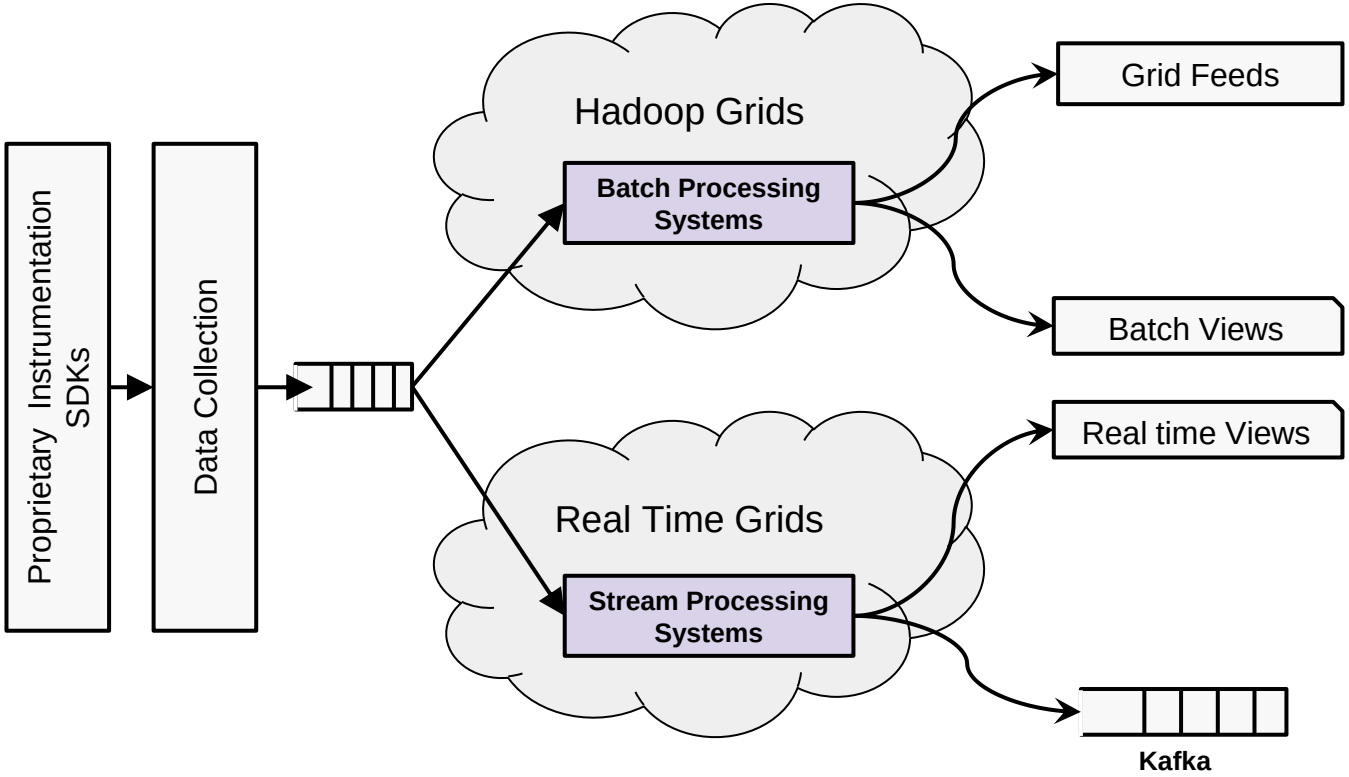
- Scale
- No downtime deploys
  - Real-time systems == “available now”
- New sources of data, “wider” data
  - How to dynamically accomodate?
- Open source
  - Balance being a user and a contributor

Thank You

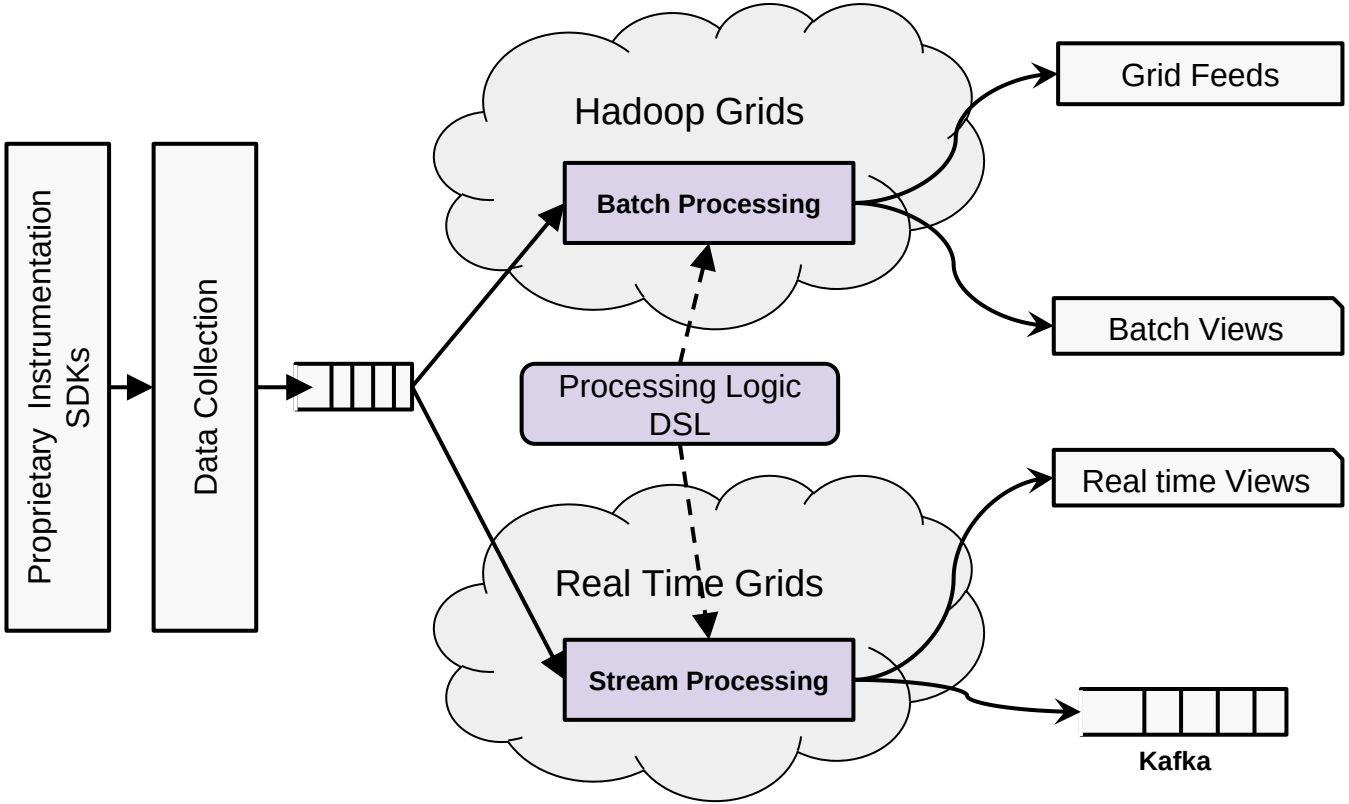
System Overview



Lambda Architecture



Lambda Architecture w/DSLs



Kappa Architecture