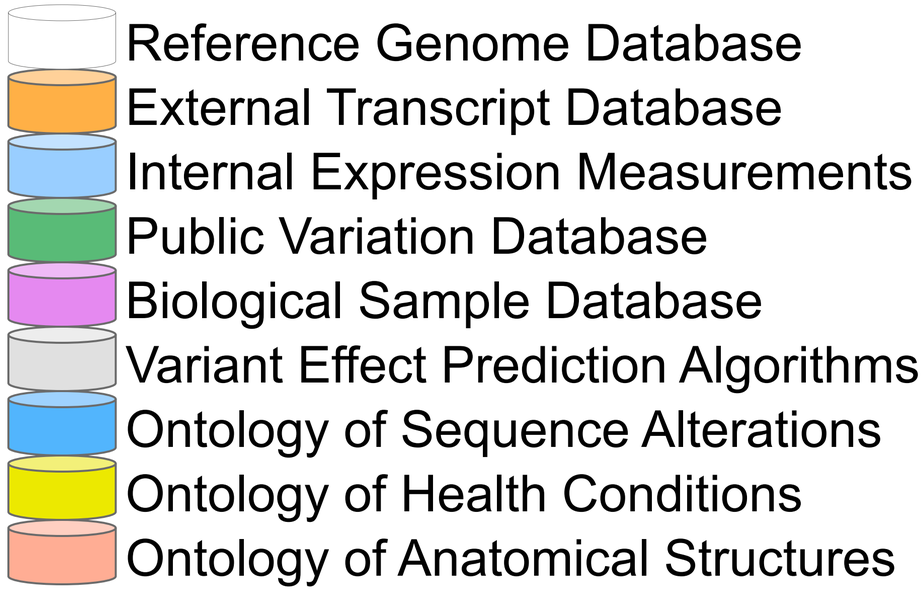
Graph-Based Genomic Integration Using Spark

David Tester

Novartis Institutes for Biomedical Research

Scope



Scope (more)

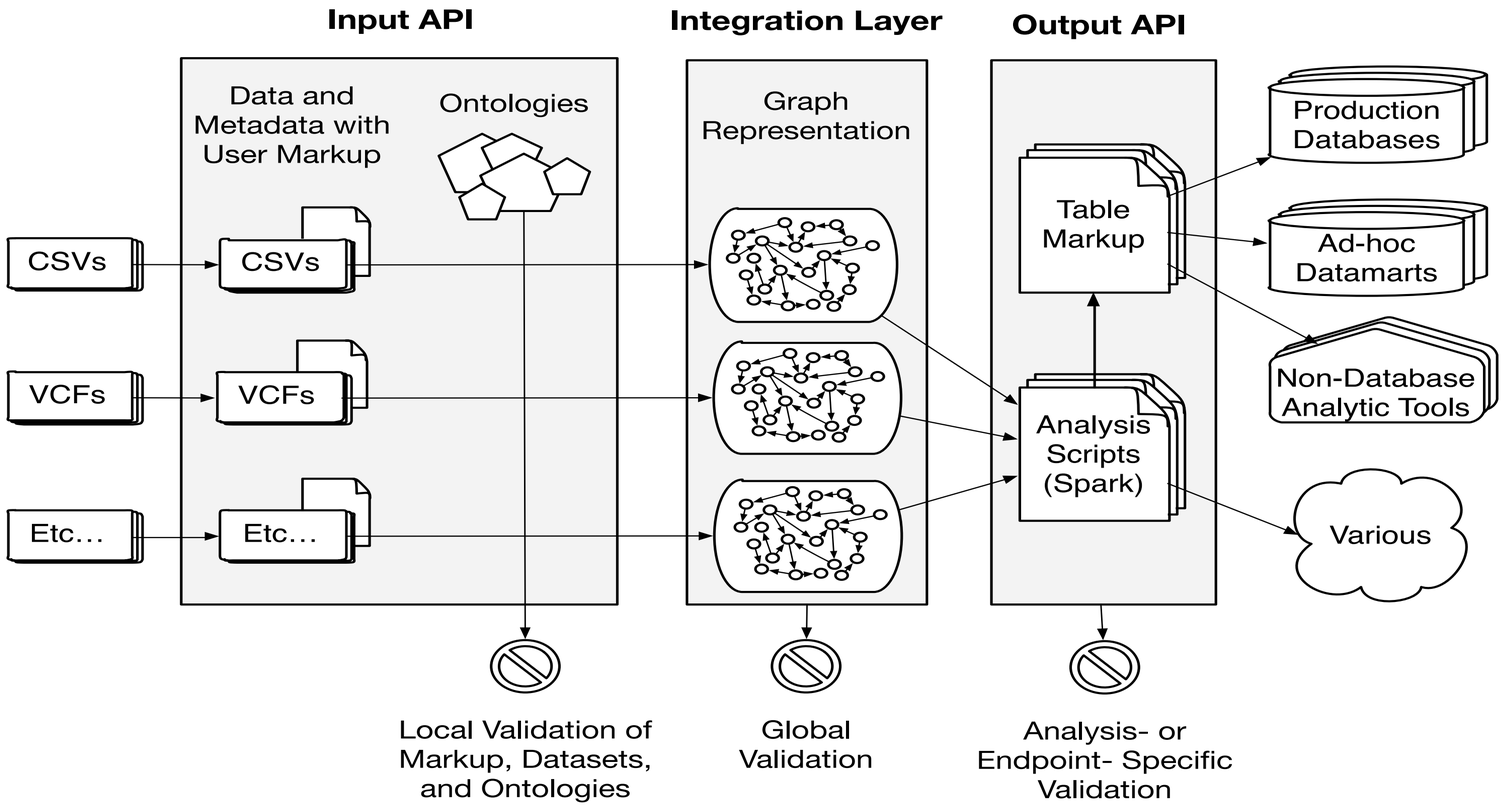
- Much, much more over time
  - New types of information (compound interactions, protein interactions, ...)
  - Confidence and Error Propagation
- Need to be extremely flexible
  - Model flexibility (our scientists disagree with each other!)
  - Analytic flexibility (technologies change!)

Data Characteristics

- Moderately sized
  - As graph: currently low trillions of edges
  - As tables: currently 100s of billions of rows
- But Growing Extremely Fast
  - Driven by logarithmic decrease in sequencing costs
- Also Extremely Messy
  - 10s-100s of internal and external sources
  - Incentive structure at odds with standardization?

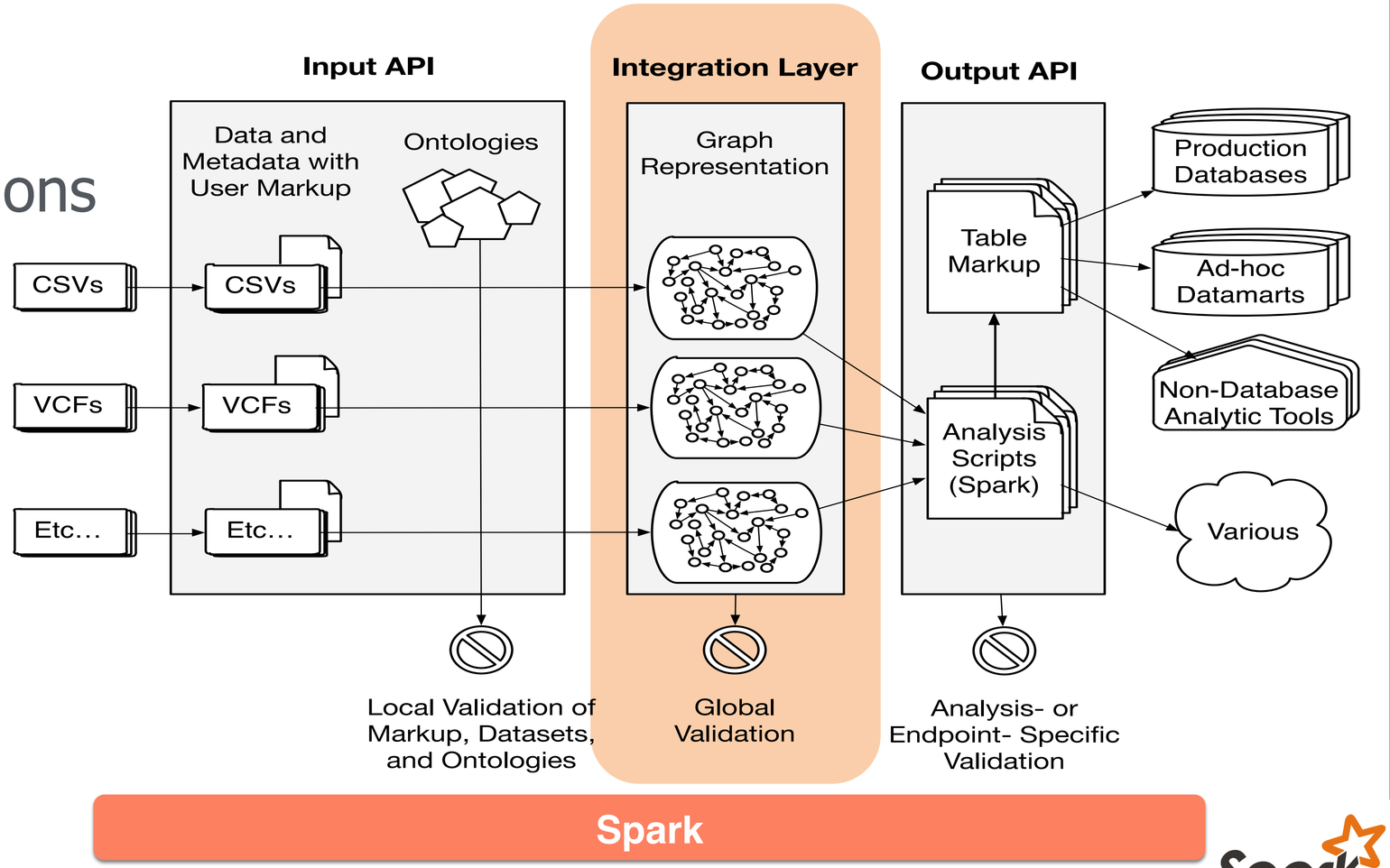
Overall Architecture

- More a Framework than a Database
- We Input:
  - Raw data and metadata
- We Add:
  - Markup (Scala DSL)
  - Ontologies
- We Archive
  - Graph representations of all datasets + markup
- We Output:
  - Vertica, SciDB, R API, Dataframes, Others…
  - Analytics API



Integration Mechanics

1. Parse
2. Group into semantic units
3. Create integration instructions
4. Join to other datasets
5. Convert to edges
6. Enrich edges (reasoning)
7. Assign IDs
8. Publish



Why Graphs?

- We really just want binary relationships
  - Example: HasDiagnosis (ThisTissue , Adenoma)
  - Flexible representation (more expressive than key/value, less rigid than tables)
  - Well-understood reasoning mechanics (transitive closures, type inference, validation, …)
  - Useful for integration
- Logically equivalent to a graph with labeled, directed edges
  - For topological and/or network analyses (e.g., shortest paths)
  - GraphX

Why not a Graph toolkit?

- Distributed graph TKs require an efficient partitioning strategy
  - Typically a function from edge or vertex to its partition
- We partition over semantic units (encoded as subgraphs)
  - Each subgraph contains the edges which describe a semantic unit
  - Has natural mapping back/forth to source files
  - Naturally expressed in Spark, awkward for GraphX
  - But GraphX is just a flatMap away when we need it!

Future

- Heavy interest in not-just-genomic data
- More reliance on Spark-based analytics
  - Parquet / Spark Dataframes / Spark SQL / Adam
- Trillion edge graph follows exponential trajectory in data size over next few years…

Acknowledgements

- Rob Anderson
- Hans Bitter
- Mark Borowsky
- Sophie Brachat
- Victor Bucor
- Rose Brannon
- Jason Calvert
- Dennis Cunningham
- John Damask
- Timothy Danford
- Anthony Dibiase
- Sean Duane
- Chris Farnham
- Nick Flower
- Laurent Gauthier
- Ajay Gourneni
- Nabil Hachem
- Victor Hong
- Mike Jones
- Jason Kondracki
- Andrew Knueven
- Igor Mendelev
- Steve Marshall
- Gregg McAllister
- Brant Peterson
- Martin Petracchi
- Brian Repko
- Erik Sassaman
- Mark Schreiber
- Ruth Seltzer
- Dave Sexton
- Anita Stout
- David Treff
- Quan Yang
- Dongmei Zuo

Many others…

We're Hiring Spark Gurus!

- Novartis Institutes for Biomedical Research - Cambridge, Massachusetts
- View our open positions at:

  http://novartis.avature.net/nibrit

NOVARTIS

NOVARTIS INSTITUTES FOR BIOMEDICAL RESEARCH