Using Impala For Variant Pipelines

Summer Elasady and Denise Mauldin

11-19-2015

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Connect!

- GitHub tutorials and PowerPoint:

https://github.com/summerela/impala_training/

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Agenda

I. What is Impala?
   I. Use Cases
   II. System Structure
   III. What's available?
II. Connecting to Impala
   I. Hue Web Interface
   II. Impala Shell
   III. R
   IV. Python
III. Creating Queries
IV. Sample Pipelines
V. Getting Help

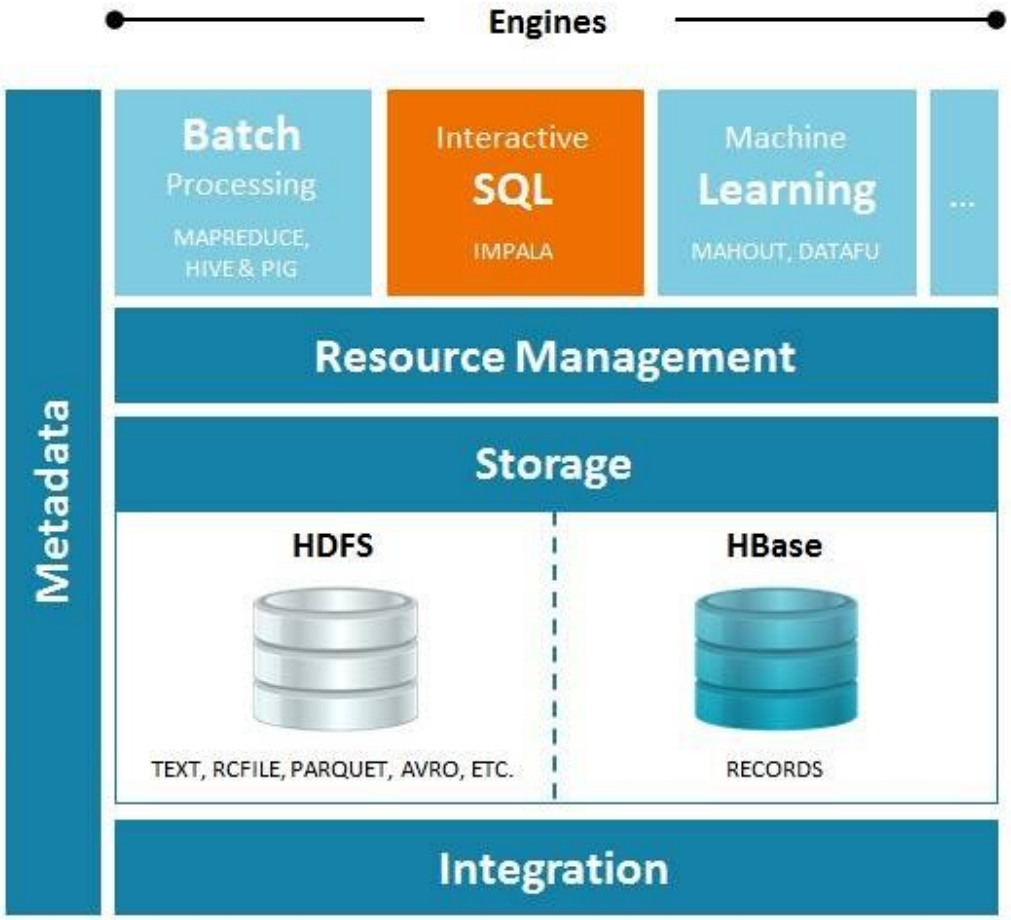
Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Use Cases: Impala in the real world

- Locate and annotate variants based on:
  - Pathogenicity
  - Gene region
  - Allele frequency
- Clarity2 Challenge
  - ACMG actionable genes
  - Pathogenic
- Candidate Script
  - Centralized annotation of candidate variants
- Newborn Screening Genes
  - Rare variants
  - Pathogenic

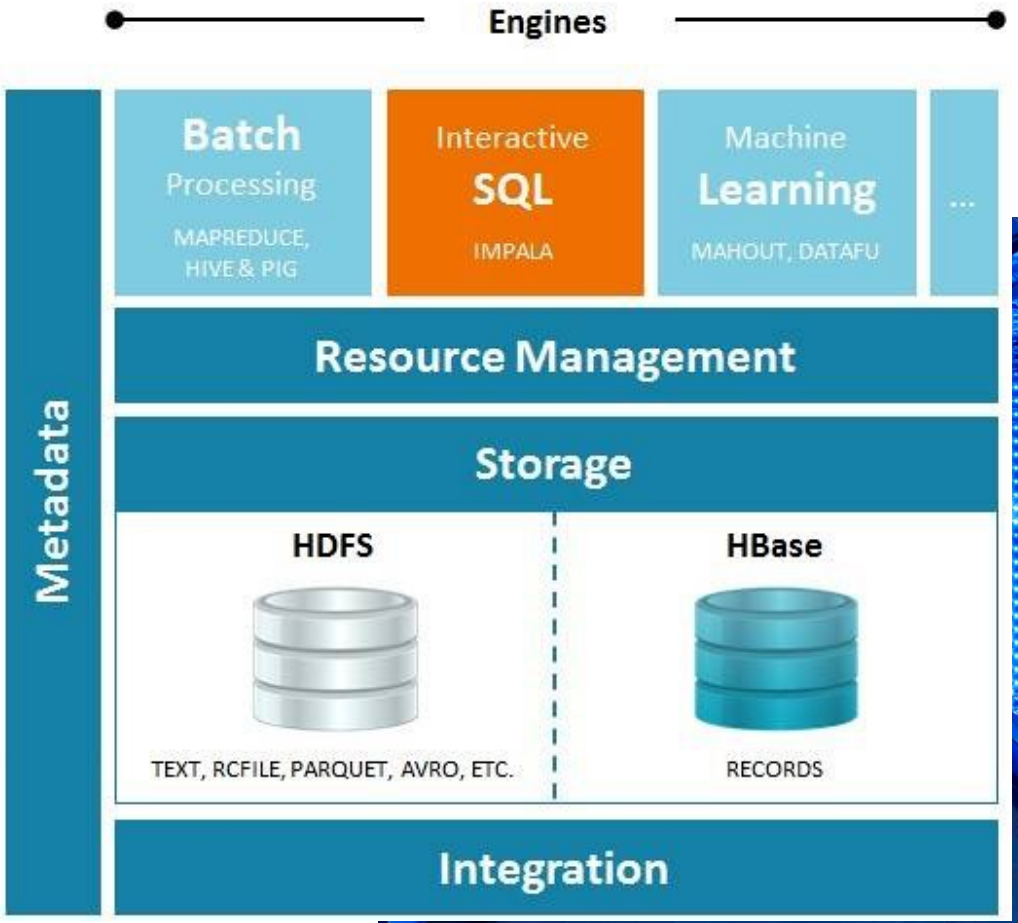
Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

What exactly is impala?



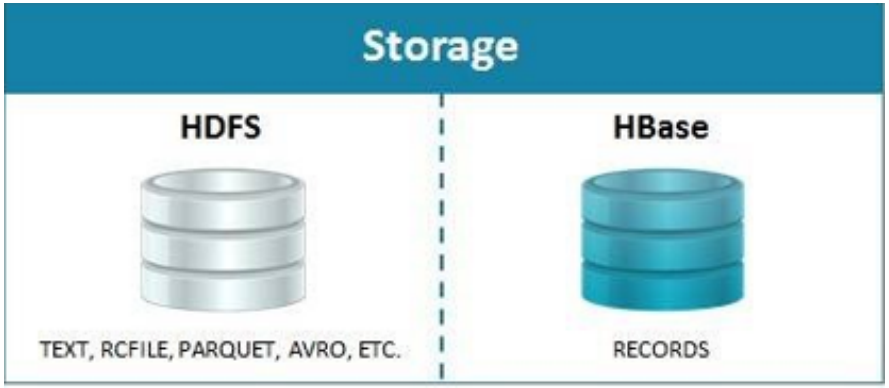
Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

What exactly is impala?



Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

What exactly is impala?

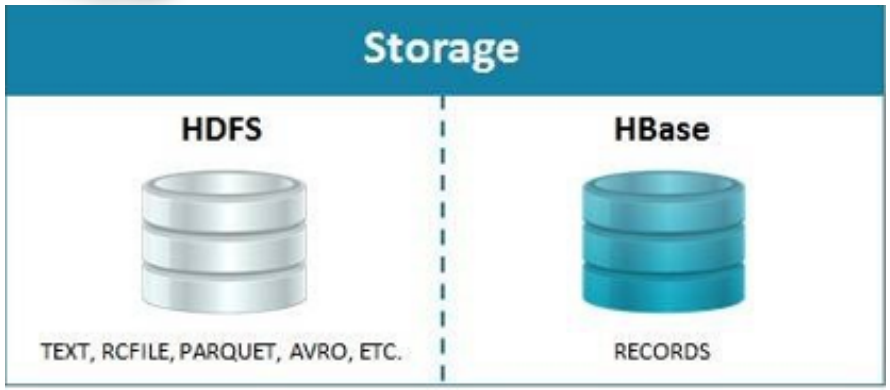


Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

What exactly is impala?

Store Files:

- Variants
- Reference DB's
- Analysis results



Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

What exactly is impala?

Raw VCF file transformed and uploaded to HDFS:

```
##fileformat=VCFv4.1
##fileDate=20150819
##source=bin/makeVCF.pl
##reference=file:///proj/famgen/resources/Kaviar-150812-ISB/bin/../tabixedRef/hg19.gz
##version=<Kaviar-150812 (hg19)>
##kaviar_url=<http://db.systemsbiology.org/kaviar>
##publication=<Glusman G, Caballero J, Mauldin DE, Hood L and Roach J (2011) KAVIAR: an accessible system for testing SNV novelty. Bi
oinformatics, doi: 10.1093/bioinformatics/btr540>
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele Count">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in data sources">
##INFO=<ID=END,Number=.,Type=Integer,Description="End position">
##INFO=<ID=DS,Number=A,Type=String,Description="Data Sources containing allele">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
1       10001   .       T       C       .       .       AF=0.0000384;AC=1;AN=26028;DS=SS6004475
1       10002   .       A       C,T     .       .       AF=0.0001153,0.0000384;AC=3,1;AN=26028;DS=HGDP00927|HGDP00998|HGDP01284,HGDP0
1029
1       10002   .       A       AT      .       .       AF=0.0000384;AC=1;AN=26028;DS=HGDP00521
1       10003   .       A       C,T     .       .       AF=0.0000384,0.0000768;AC=1,2;AN=26028;DS=HGDP01284,HGDP00521|HGDP00927
1       10004   .       C       A       .       .       AF=0.0000384;AC=1;AN=26028;DS=HGDP01284
1       10018   .       C       T       .       .       AF=0.0000384;AC=1;AN=26028;DS=HGDP00998
1       10019   rs775809821     TA      T       .       .       AF=0.0000384;AC=1;AN=26028;END=10020;DS=Malay
```

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

What exactly is impala?

Read as a table by impala:

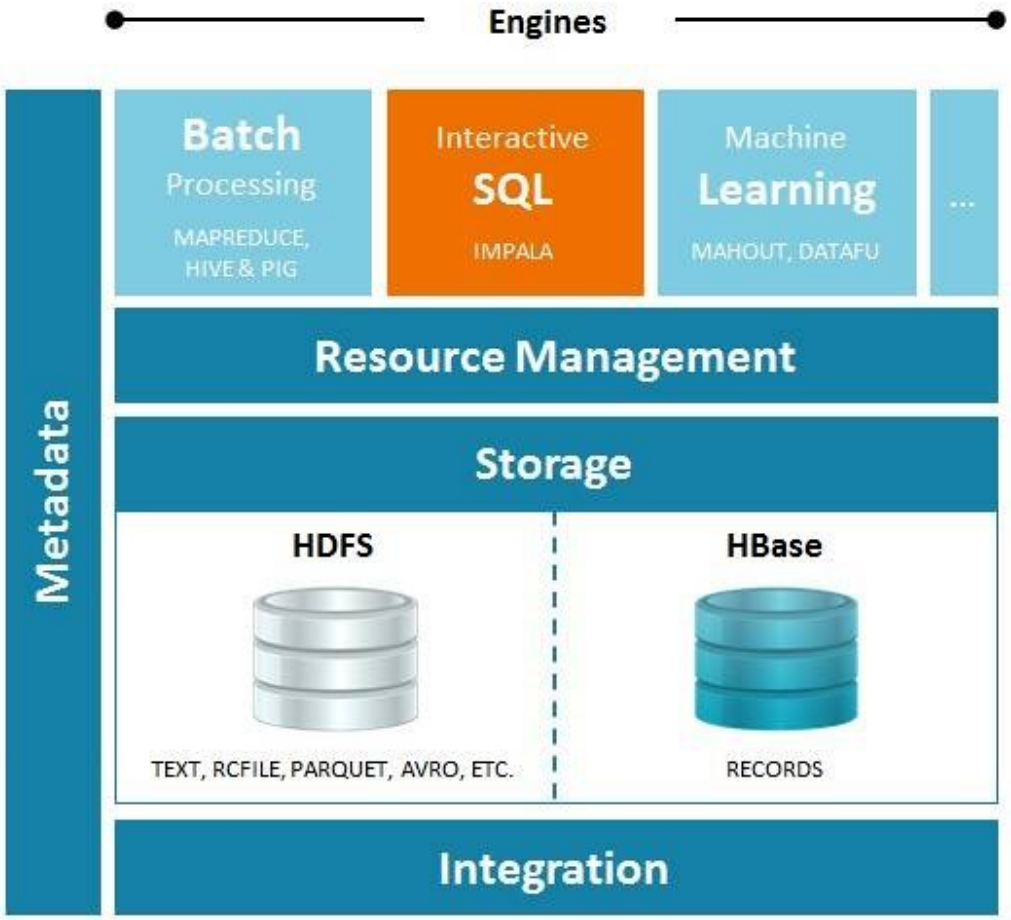
Data sample for kaviar

View in Metastore Browser

chrom	pos	stop	rs_id	ref	alt	qual	filter	allele_freq	allele_cnt	allele_num	sources
10	98917931	None	rs75578462	C	T	None	None	0.053711399436	1398	26028	ADNI|GMIAK9|GS000009930|GS0000117
10	98917981	None	rs545362632	C	T	None	None	0.000230499994359	6	26028	phase3-MSL
10	98917985	None	rs17112469	T	C	None	None	0.106577500701	2774	26028	!Gubi|ADNI|Desmond Tutu|GMIAK6|GMIA
10	98918018	None	rs746271737	G	A	None	None	3.84000013582e-05	1	26028	UK10K
10	98918073	None	rs756450057	T	C	None	None	3.84000013582e-05	1	26028	UK10K

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

What exactly is impala?



Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

What exactly is impala?

What exactly is impala?

SQL – Structured query language

SELECT columns
FROM database.table_name
WHERE parameters

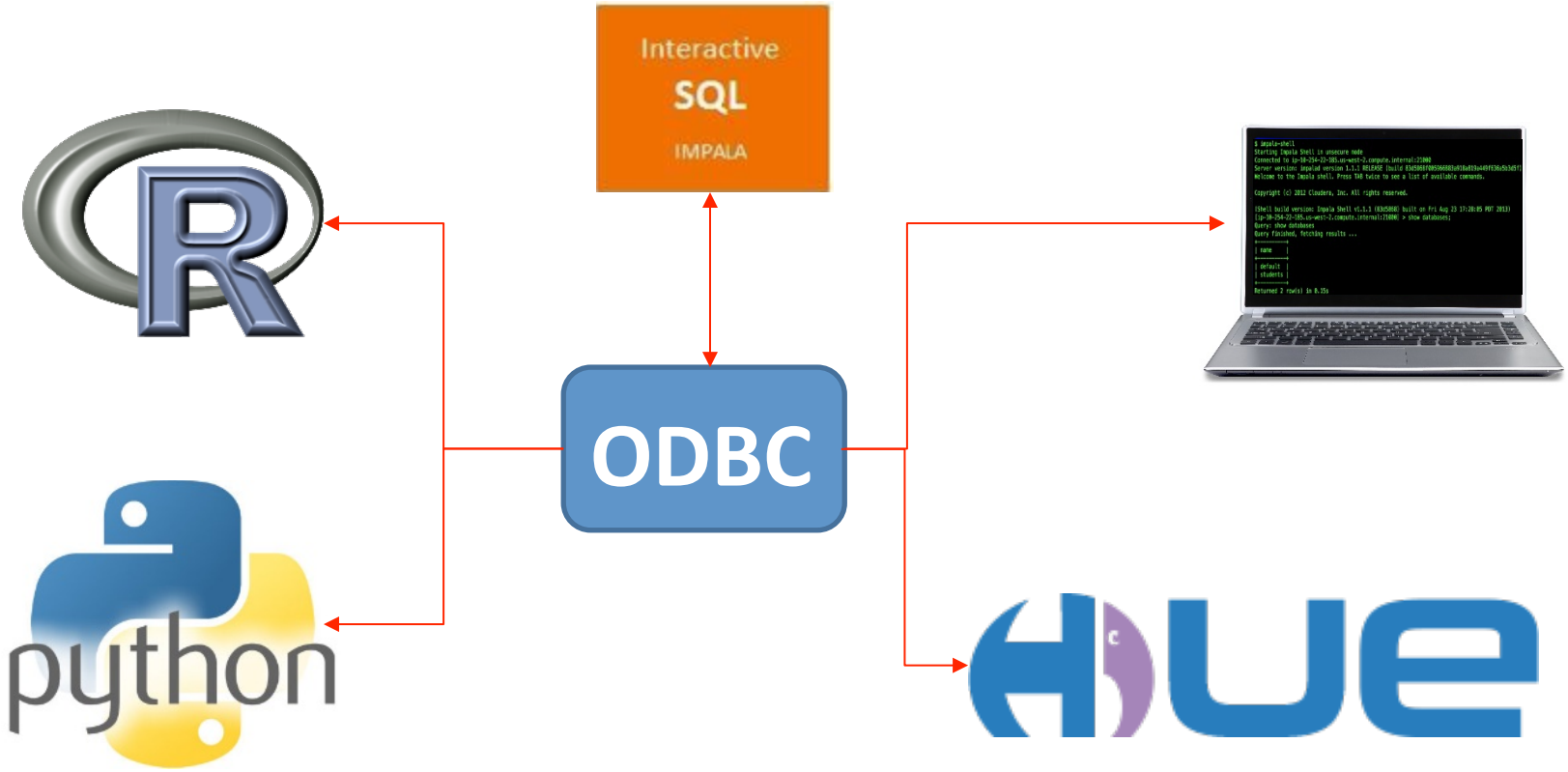
What exactly is impala?

Interact with impala to:

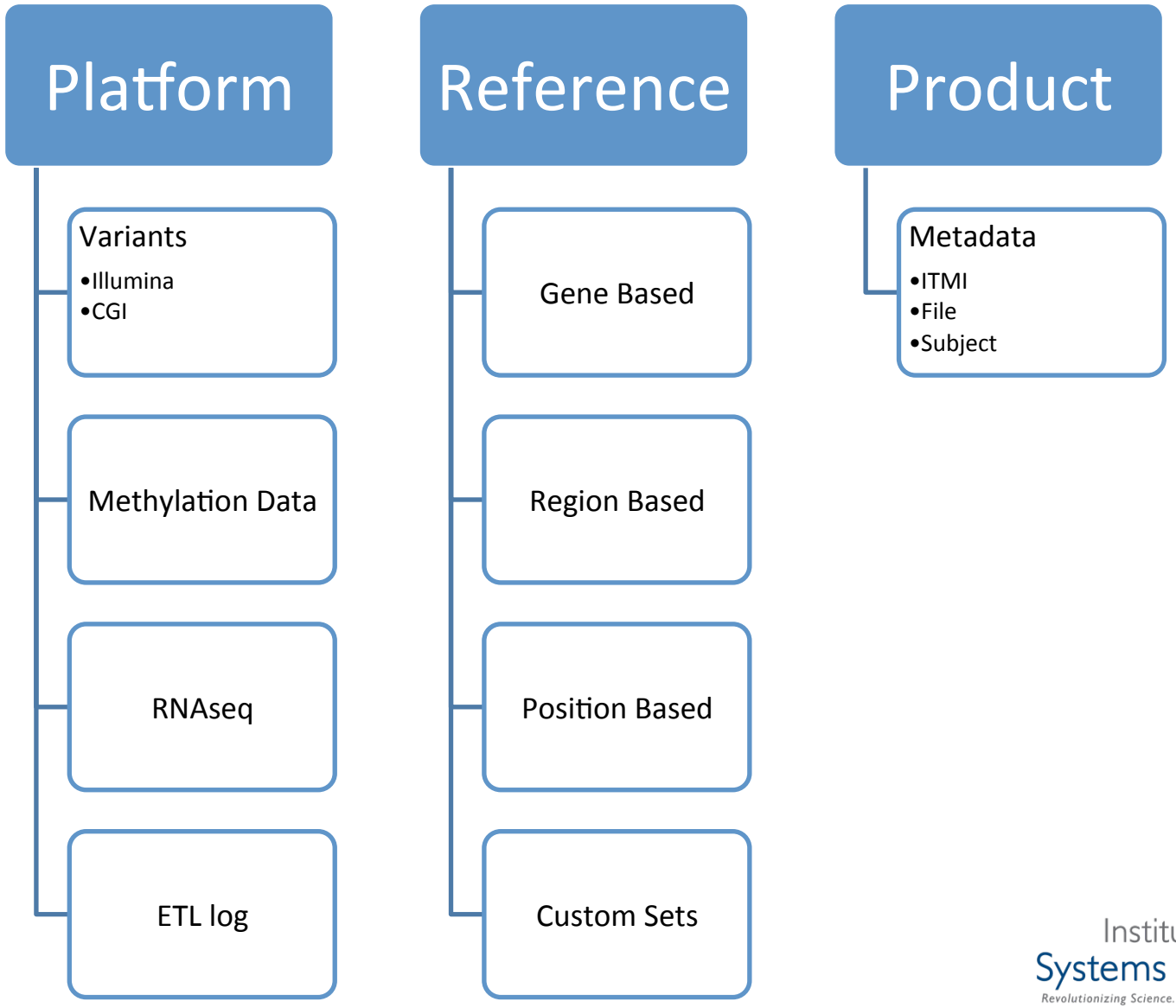
- Filter
- Locate
- Annotate

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

What exactly is impala?



Database Structure



Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

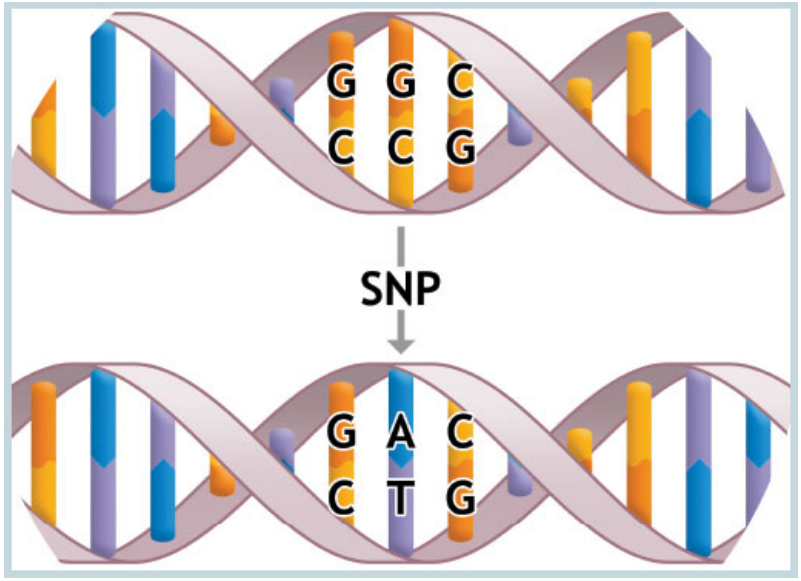
Benefits

- Consistency
  - Normalized Data
  - 1-based coordinates
  - Universal Column Names
- Speed
  - Quickly run queries across thousands of genomes

What's Available: Variants

Complete genomics

illumina®



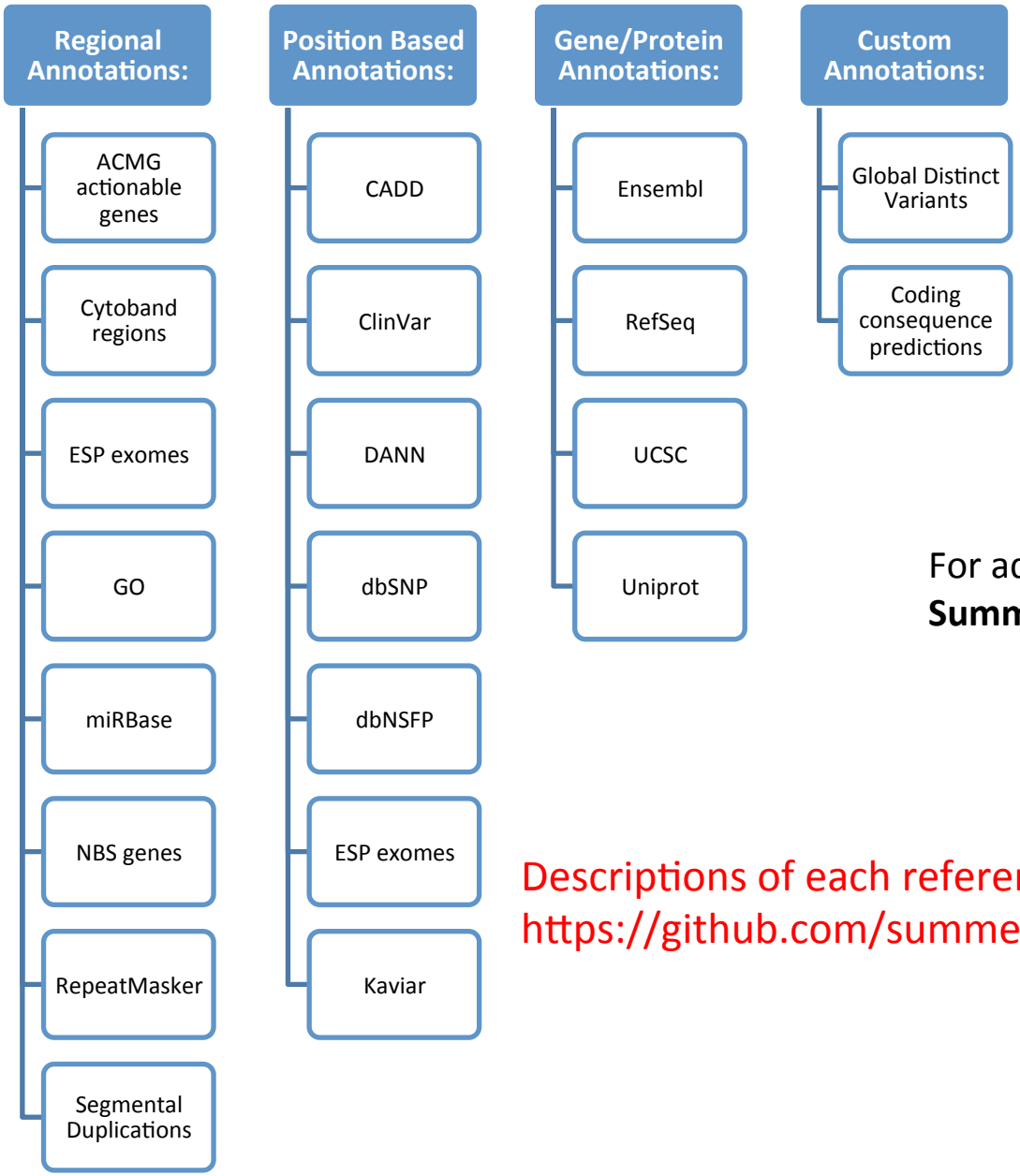
101

102

103

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Reference Sources



For additional annotations email:
Summer.Elasady@systemsbiology.org

Descriptions of each reference:
https://github.com/summerela/external-reference-data-catalog

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Accessing Impala

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Accessing Impala

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Accessing Impala

cloudera® IMPALA

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Accessing Impala

cloudera® IMPALA

Pros:

- Easy, web based, quick
- Easy to do query testing and EXPLAIN plan

HUE

Cons:

- Cuts off results inconsistently at 100,000 rows
- AJAX connection issues:
  - Connection drops
  - Canceling queries

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Accessing Impala

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Accessing Impala

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Accessing Impala

cloudera® IMPALA

Pros:

- Fast
- Efficient

Cons:

- Less intuitive and friendly interface

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

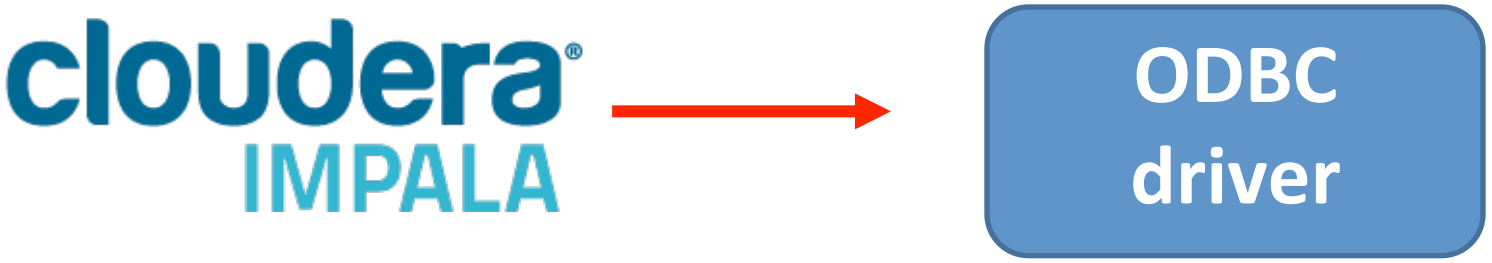
Accessing Impala

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Accessing Impala

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Accessing Impala



Accessing Impala

cloudera IMPALA → **ODBC driver**

RODBC

Impyla
Ibis

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Hue Interface

ISB Impala Login:

- Login to your impala web interface

Hue Tutorial:

https://github.com/summerela/impala_training/blob/master/using_hue.pdf

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Connecting with impala-shell

- Tutorial:

  https://github.com/summerela/impala_training/blob/master/impala_shell.ipynb

Connecting with Python

- Python:

  https://github.com/summerela/impala_training

  – Click on the ‘launch binder’ icon

  – Click on connect_python.ipynb

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Connecting with R

- R:
  - Launch R:
    - Launch your R web interface
  - Tutorial:

    https://github.com/summerela/impala_training/blob/master/connect_with_R.md
  - Script:
    https://github.com/summerela/impala_training/blob/master/connect_R.R

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Building Queries

- Choose your interface
- Follow along (and copy/paste queries):

  https://github.com/summerela/impala_training/blob/master/building_queries.md

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Basic Pipeline

1. Connect to impala
2. Input genes of interest
3. Input subject id's of interest
4. Annotate
5. Filter
6. Export results

https://github.com/summerela/impala_training/blob/master/variant_pipeline_python.ipynb

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.