Rancho & Pfizer working together for the tranSMART Community

PM: Tatiana Khasanova  
Admin: Julie Bryant  
April, 2015
Customer wanted multiple data types in tranSMART

- Customer chose to have Pfizer GWAS data in tranSMART 1.2
- Rancho could not get the script to work and called Pfizer
- Pfizer very responsive
- We worked together for a solution
- New scripts, documentation in Github
Tested GWAS package for Oracle

- Data dumps for annotation tables
- Migration SQL scripts for stored procedures
  - Pfizer rolled out scheme of DB but not migration script so we could not apply to existing DB
  - Rancho created migration script to make it work easier-call scripts sequentially
- Indexes for better performance in GWAVA
  - Uses GWAVA to read GWAS data, was slow-made changes to increase speed significantly
- Pfizer Updated Kettle scripts
  - Rancho loaded in Github with package
- Configurations in tranSMART
  - Tweaks to improve for environment with SOLR server-full text search server-grabs any content PDF word etc can configure to search everywhere. Rancho grabs GWAS info meta data. DB working with raw text. Search fast.
GWAS data flow

1. Data loading from files with Kettle scripts into buffer schema
2. In database data transfer from buffer storage into indexed prod schema
3. Meta info indexing by SOLR
4. Meta info retrieval
5. Variants data retrieval

Very compute consuming and lengthy operation.
Room for optimization:
- additional indexing
- structure refactoring for read operations
- db configuration

Data consumers
GWAS module
GUAVA

Loading workflow
Results

- Instruction tested on public clear build TM >= v1.2.2 (Oracle 11g)
- MAGIC dataset is a sample for fast test of concept
- Approach has been used on two big pharma GWAS datasets (10 GB)
  - Loading time took less than 2 days
- Gives community a sense of how long it should take to compare how they are doing
Future plans

- Test on PostgreSQL version
- Optimize data retrieval operation for GUAVA
- Fix issues with GWAS UI
- All changes posted in Github
Thank you for your time

GWAS ETL package link:
Thank you

Merck
Meiping
Scott

Pfizer
Jay
Haiyan
Hugo

Rancho
Alex
Tania