

How we use Apache Pig

Stefan Groschupf

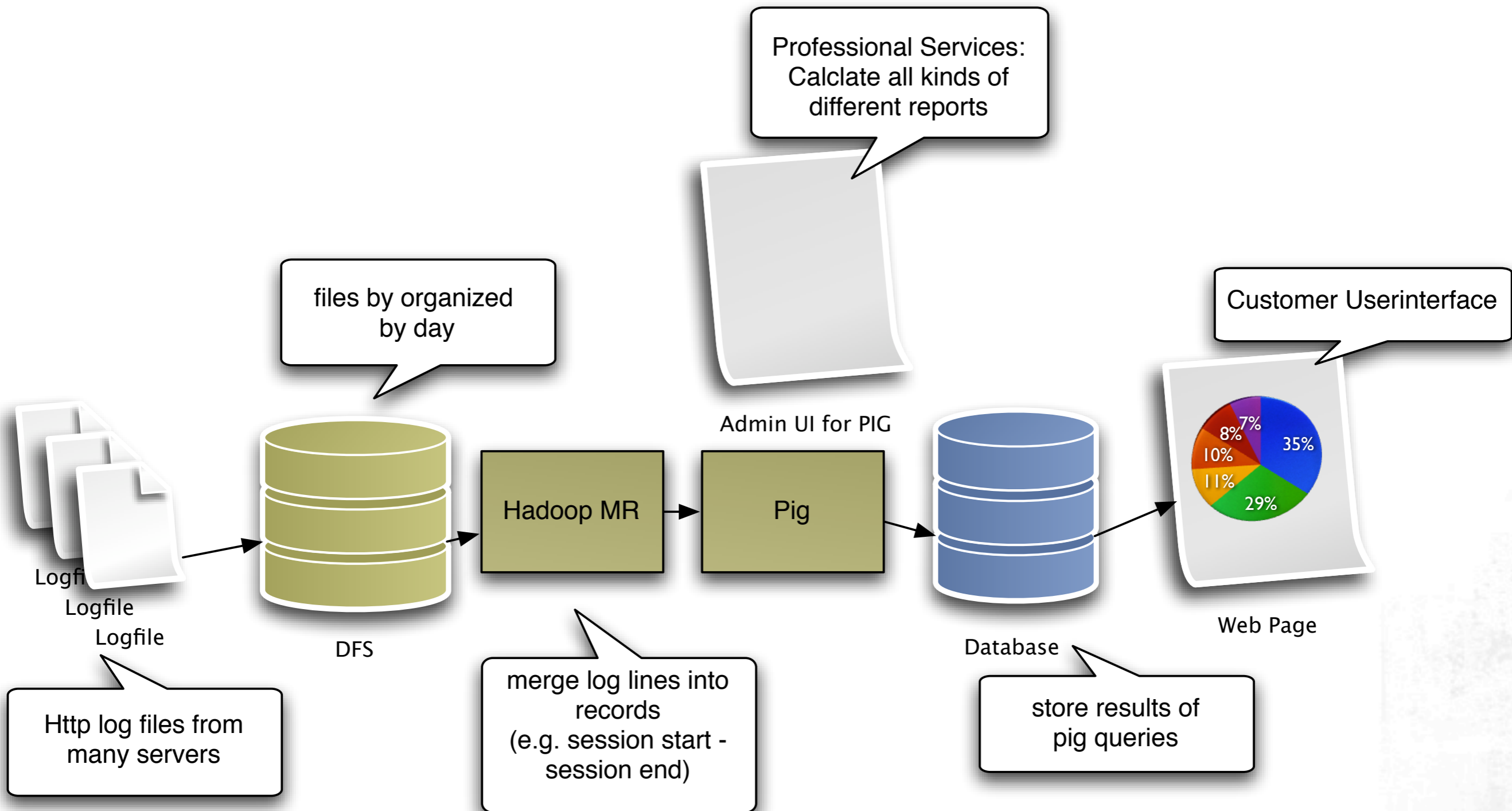
sg{at}101tec.com

101tec Inc.

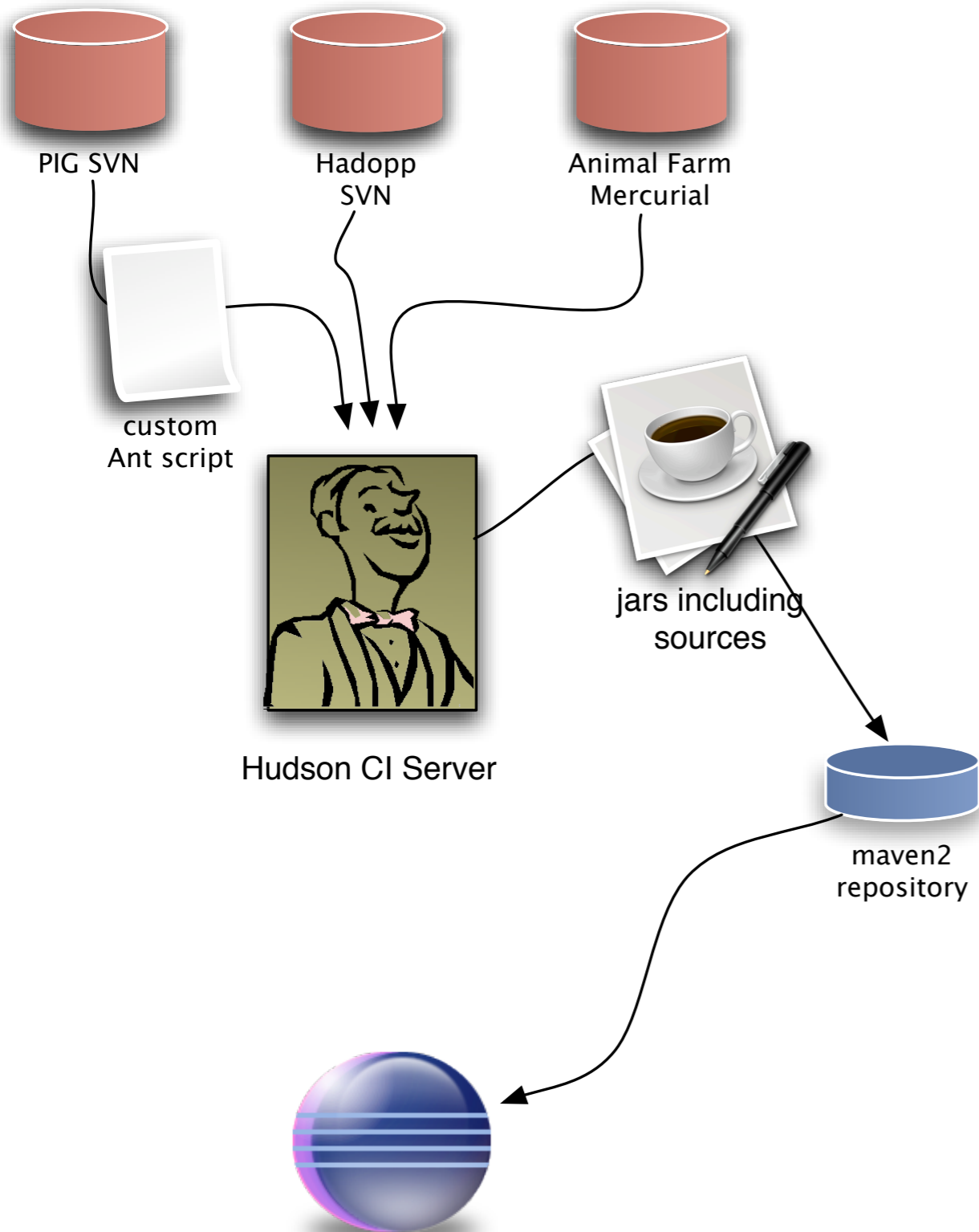
Menlo Park, CA

- **System overview**
- **Integration into our build**
- **Entity compiling in Hadoop**
- **Hadoop >> Pig >> Oracle**
- **Problems**

System overview “animal farm”

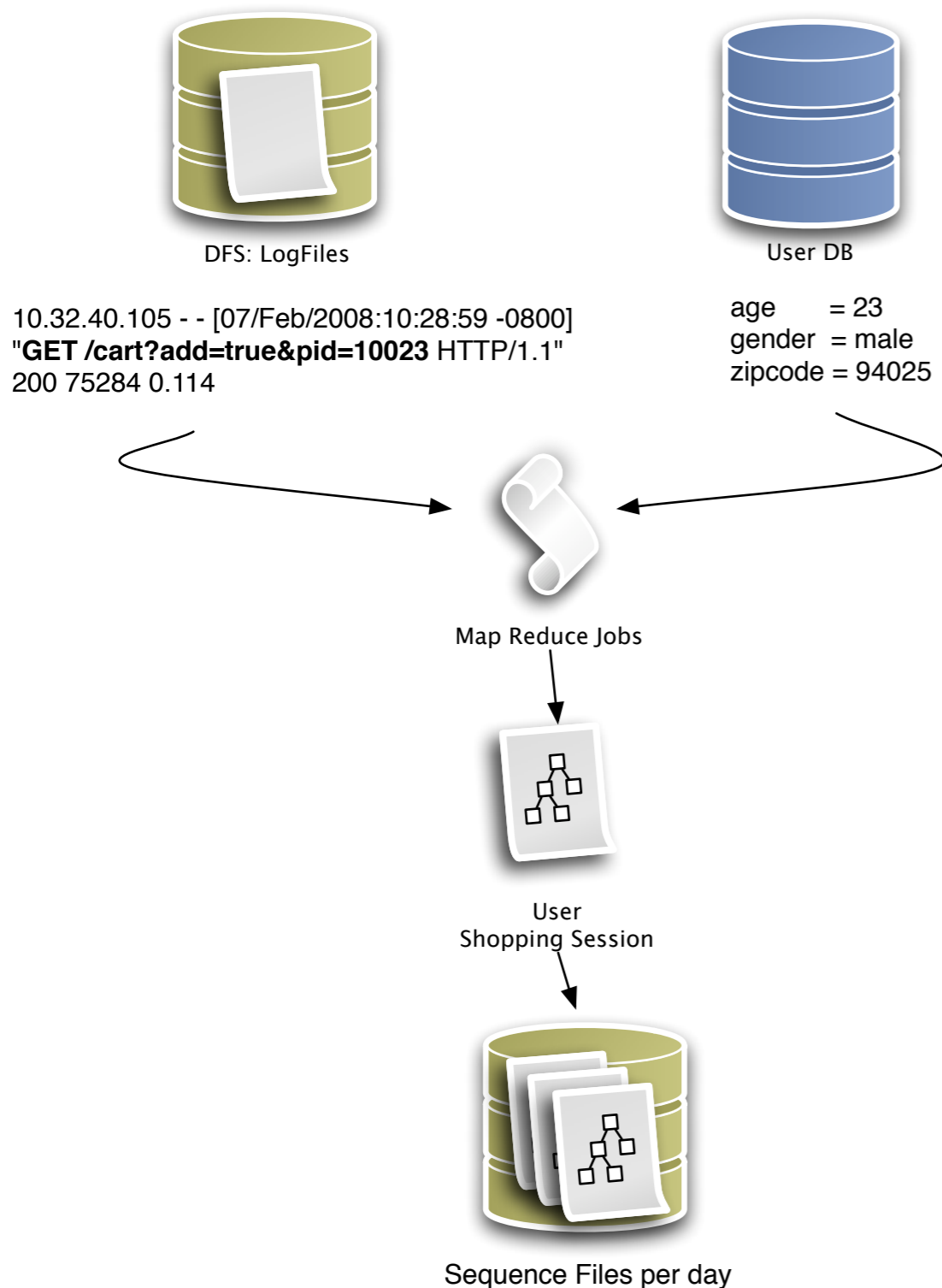


Integration into our build



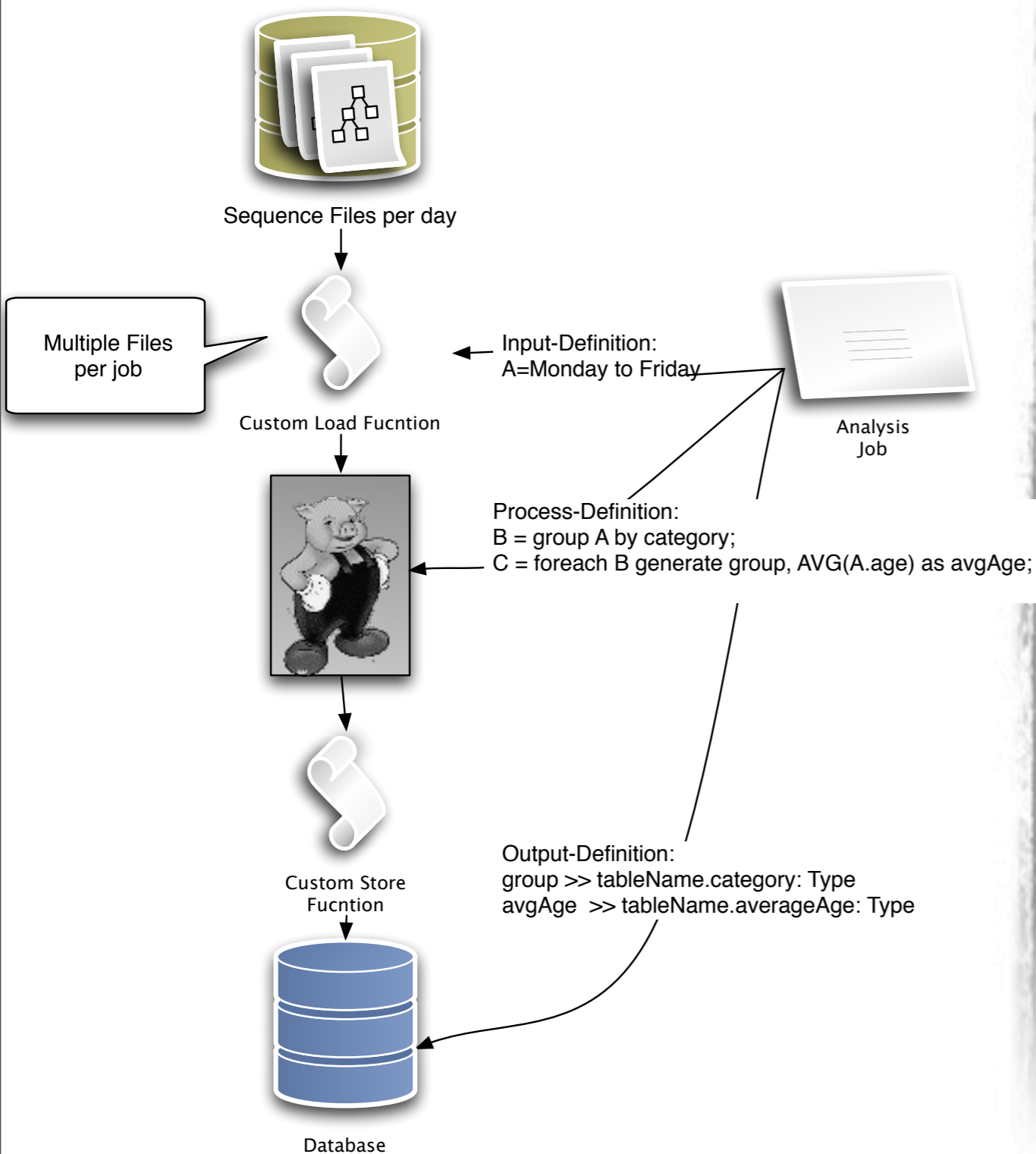
- **CI build for all components**
- **Animal farm: maven2 build**
- **Pig build: ant**
- **Plus custom script to add sources to jar.**
- **Deployed to company maven repository**

Entity compiling in Hadoop



- **Compile events to entity**
- **Accumulate user meta data**
- **Sequence of Map Reduce Jobs**
- **Store result in tmp sequence file**

Hadoop >> Pig >> Database



- Custom load function into Pig
- User can define attributes that are used for Pig Queries
- Entity(types) to Tuple (strings)
- Custom Store Function
- Automatically generate table schema
- Tuples > batch inserts into DB

Problems, we did run into

- **No official jar repository, no source jar**
 - **Custom build**
- **Pig jar contains third party classes (junit, hadoop, etc)**
- **Centralized logging**
 - **log4j support? Pig has no logging yet. (PIG-83)**
- **Exception handling & System.exit**
- **More flexible LoadFunc & StoreFunc**
 - **Configurable**
 - **Files need to exist**
 - **Multiple Files**

sg{at}101tec.com.

