Dataiku Flow and dctc

Data pipelines made easy

- Berlin Buzzwords 2013

Dataiku

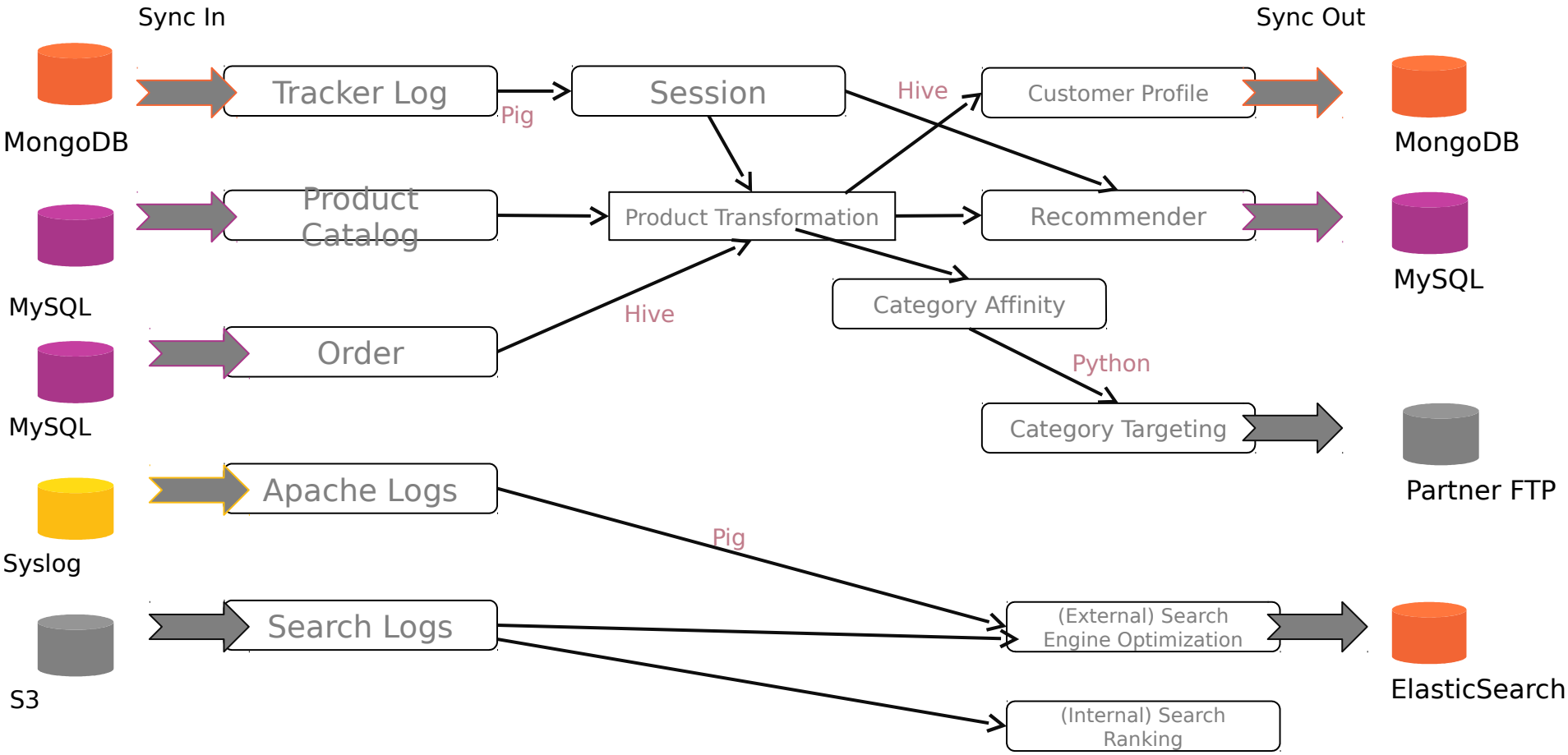
About me

Clément Stenac <clement.stenac@dataiku.com>
@ClementStenac

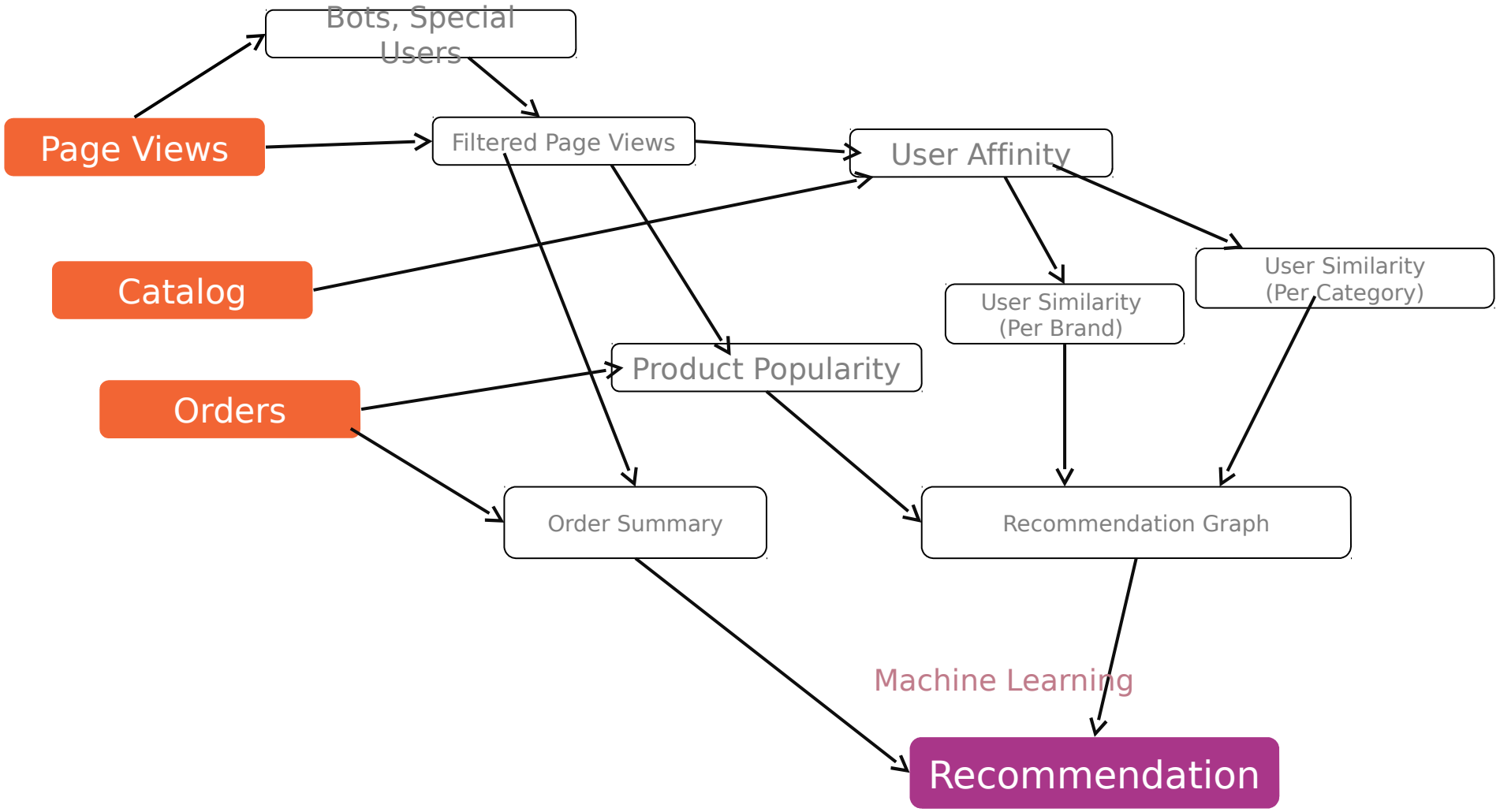
- CTO @ Dataiku
- Head of product R&D @ Exalead
  (Search Engine Technology)
- OSS developer @ VLC, Debian and OpenStreetMap

- The hard life of a Data Scientist
- Dataiku Flow
- DCTC
- Lunch !

Follow the Flow

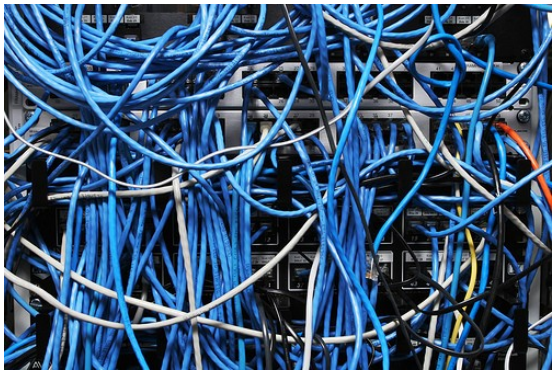


Zooming more



Real-life data pipelines

- Many tasks and tools
- Dozens of stage, evolves daily
- Exceptional situations are the norm
- Many pains
  - Shared schemas
  - Efficient incremental synchronization and computation
  - Data is bad



An evolution similar to build

Dataiku

- 1970 Shell scripts
- 1977 Makefile
- 1980 Makedeps
- 1999 SCons/CMake
- 2001 Maven

- Better dependencies
- Higher-level tasks

- … Shell Scripts
- 2008 HaMake
- 2009 Oozie
- ETLS, …
- Next ?

- The hard life of a Data Scientist
- Dataiku Flow
- DCTC
- Lunch !

Introduction to Flow

Dataiku

Dataiku Flow is a data-driven orchestration framework for complex data pipelines

- Manage data, not steps and taks
- Simplify common maintainance situations
  - Data rebuilds
  - Processing steps update
- Handle real day-to-day pains
  - Data validity checks
  - Transfers between systems

Concepts: Dataset

- Like a table : contains records, with a schema
- Can be partitioned
  - Time partitioning (by day, by hour, …)
  - « Value » partitioning (by country, by partner, …)
- Various backends
  - Filesystem
  - HDFS
  - ElasticSearch
  - SQL
  - NoSQL (MongoDB, …)
  - Cloud Storages

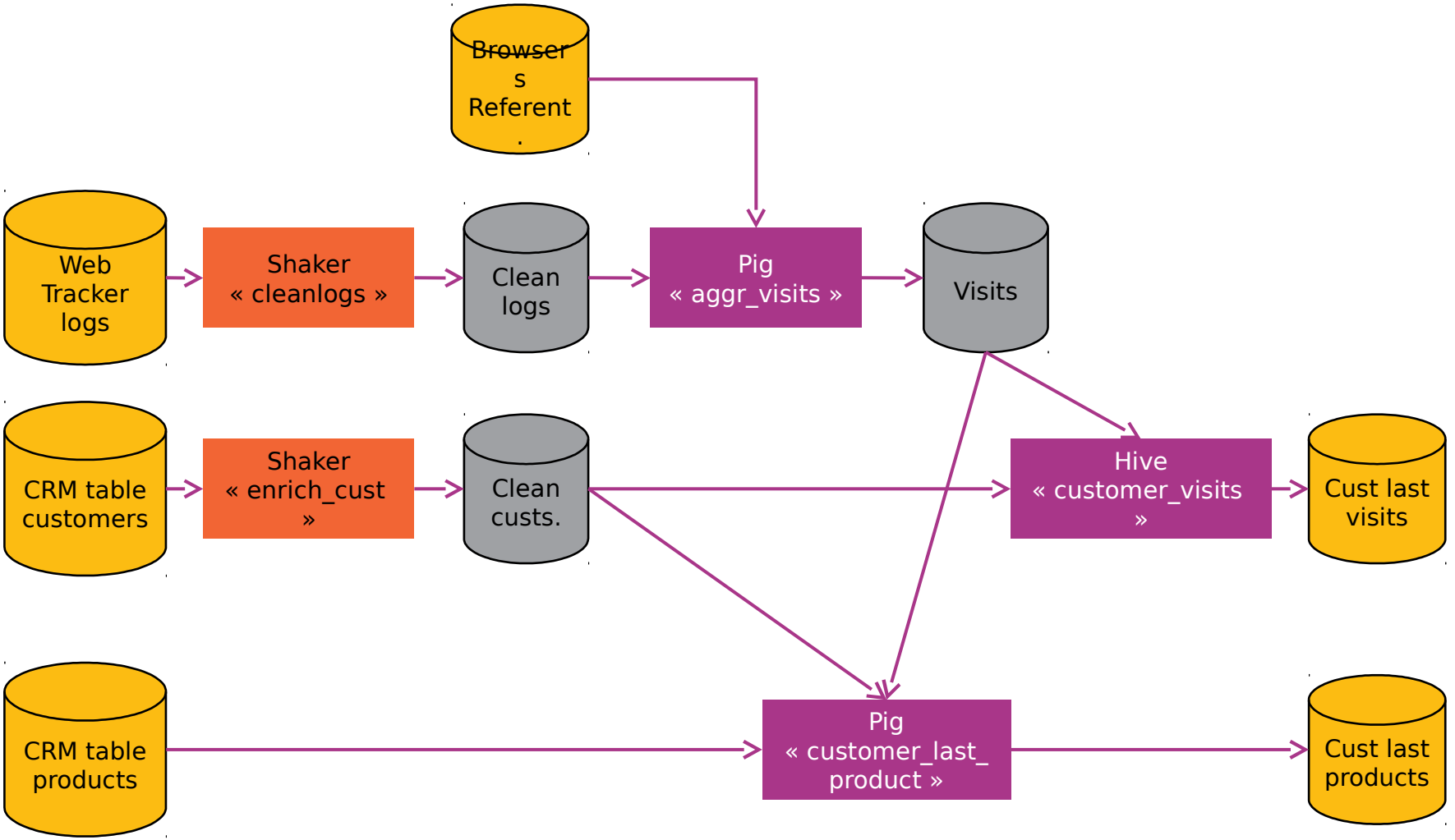
Concepts: Task

- Has input datasets and output datasets

- Declares dependencies from input to output

- Built-in tasks with strong integration
  - Pig
  - Hive
  - Python Pandas & SciKit
  - Data transfers
- Customizable tasks
  - Shell script, Java, ...

Introduction to Flow
A sample Flow



Data-oriented

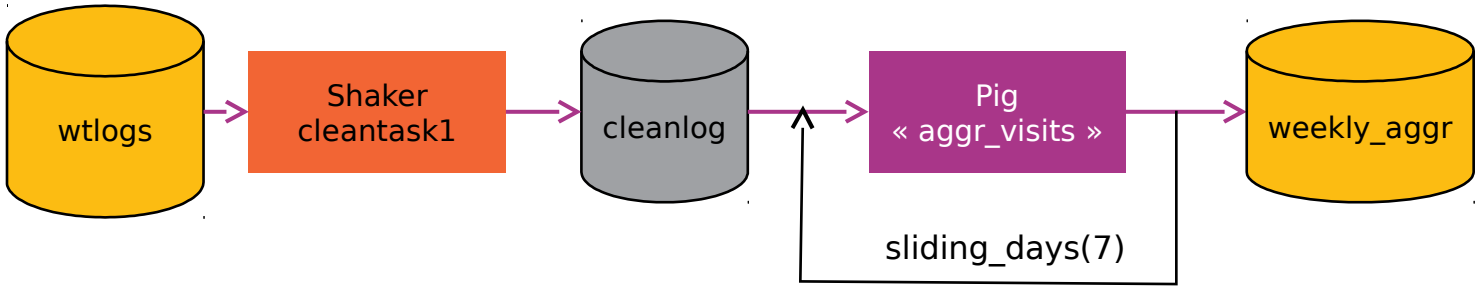
Flow is **data-oriented**

- Don’t ask « Run task A and then task B »
- Don’t even ask « Run all tasks that depend from task A »

- Ask « Do what’s needed so that my aggregated customers data for 2013/01/25 is up to date »

- Flow manages dependencies between datasets, through tasks
- You don’t execute tasks, you compute or refresh datasets

Partition-level dependencies



- "wtlogs" and "cleanlog" are day-partitioned
- "weekly_aggr" needs the previous 7 days of clean logs

- "sliding days" partition-level dependency

- "Compute weekly_aggr for 2012-01-25"
  - Automatically computes the required 7 partitions
  - For each partition, check if cleanlog is up-to-date wrt. the wtlogs partition
  - Perform cleantask1 in parallel for all missing / stale days
  - Perform aggr_visits with the 7 partitions as input

Automatic parallelism

- Flow computes the global DAG of required activities
- Compute activities that can take place in parallel
- Previous example: 8 activities
  - 7 can be parallelized
  - 1 requires the other 7 first

- Manages running activities
- Starts new activities based on available resources

Schema and data validity checks

- Datasets have a schema, available in all tools
- Advanced verification of computed data
  - "Check that output is not empty"
  - "Check that this custom query returns between X and Y records"
  - "Check that this specific record is found in output"
  - "Check that number of computed records for day B is no more than 40% different than day A"
- Automatic tests for data pipelines

Integrated in Hadoop, open beyond

- Native knowledge of Pig and Hive formats
- Schema-aware loaders and storages

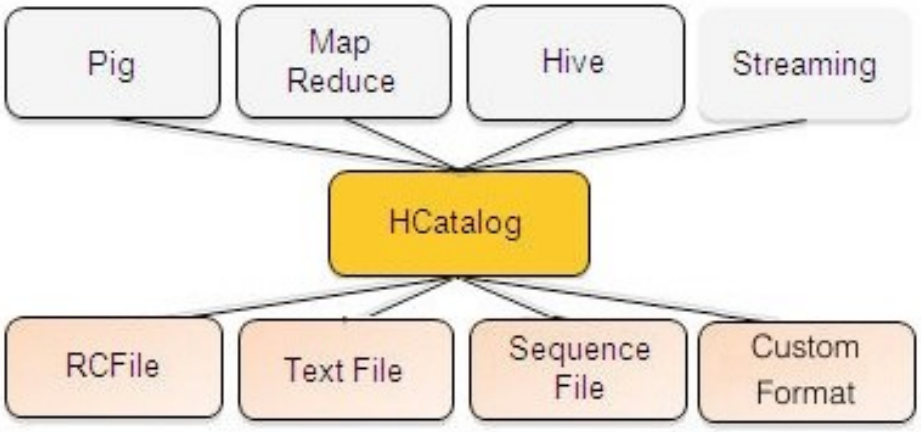
- A great ecosystem, but not omnipotent
  - Not everything requires Hadoop's strong points

- Hadoop = first-class citizen of Flow, but not the only one

- Native integration of SQL capabilities
- Automatic incremental synchronization to/from MongoDB, Vertica, ElasticSearch, ...
- Custom tasks

What about Oozie and Hcatalog ?

Dataiku



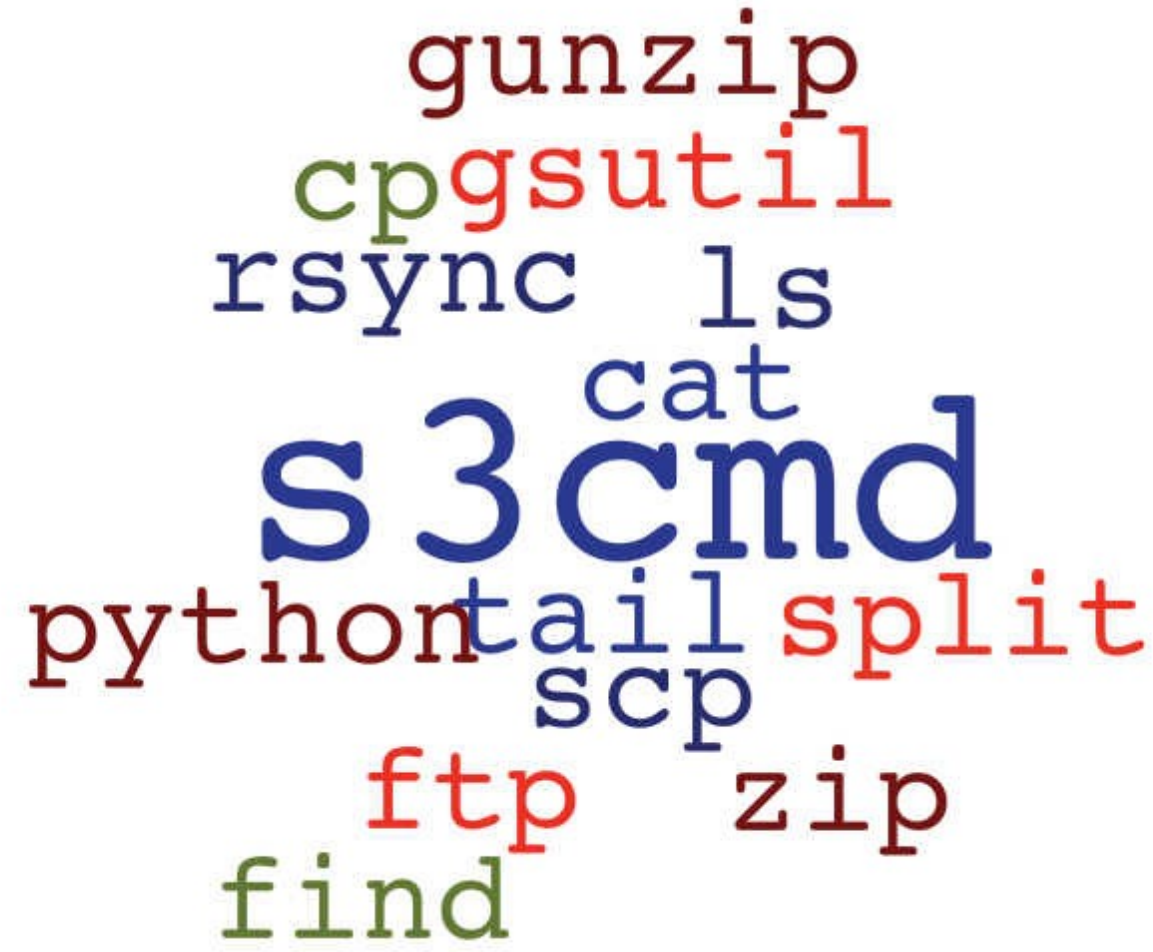
Are we there yet ?

- Engine and core tasks are working
- Under active development for first betas
- Get more info and stay informed
  http://flowbeta.dataiku.com

And while you wait, another thing

Ever been annoyed by data transfers ?

Feel the pain



- The hard life of a Data Scientist
- Dataiku Flow
- DCTC
- Lunch !

DCTC : Cloud data manipulation

- Extract from the core of Flow
- Manipulate files across filesystems

```
# List the files and folders in a S3 bucket
% dctc ls s3://my-bucket

# Synchronize incrementally from GCS to local folder
% dctc sync gs://my-bucket/my-path target-directory

# Copy from GCS to HDFS, compress to .gz on the fly
# (decompress handled too)
% dctc cp -R -c gs://my-bucket/my-path hdfs:///data/input

# Dispatch the lines of a file to 8 files on S3, gzip-compressed
% dctc dispatch input s3://bucket/target -f random -nf 8 -c
```

DCTC : More examples

```
# cat from anywhere
% dctc cat ftp://account@:/pub/data/data.csv

# Multi-account aware
% dctc sync s3://account1@path s3://account2@other_path

# Edit a remote file (with $EDITOR)
% dctc edit ssh://account@:myfile.txt

# Transparently unzip
% dctc

# Head/tail from the cloud
% dctc tail s3://bucket/huge-log.csv
```

Try it now

Dataiku

Fork me on GitHub

http://dctc.io

- Self-contained binary for Linux, OS X, Windows

- Amazon S3
- Google Cloud Storage
- FTP
- HTTP
- SSH
- HDFS (through local install)

Questions ?

Florian Douetteau
Chief Executive Officer
florian.douetteau@dataiku.com
+33 6 70 56 88 97
@fdouetteau

Marc Batty
Chief Customer Officer
marc.batty@dataiku.com
+33 6 45 65 67 04
@battymarc

Thomas Cabrol
Chief Data Scientist
thomas.cabrol@dataiku.com
+33 7 86 42 62 81
@ThomasCabrol

Clément Stenac
Chief Technical Officer
clement.stenac@dataiku.com
+33 6 28 06 79 04
@ClementStenac

Dataiku

Dataiku