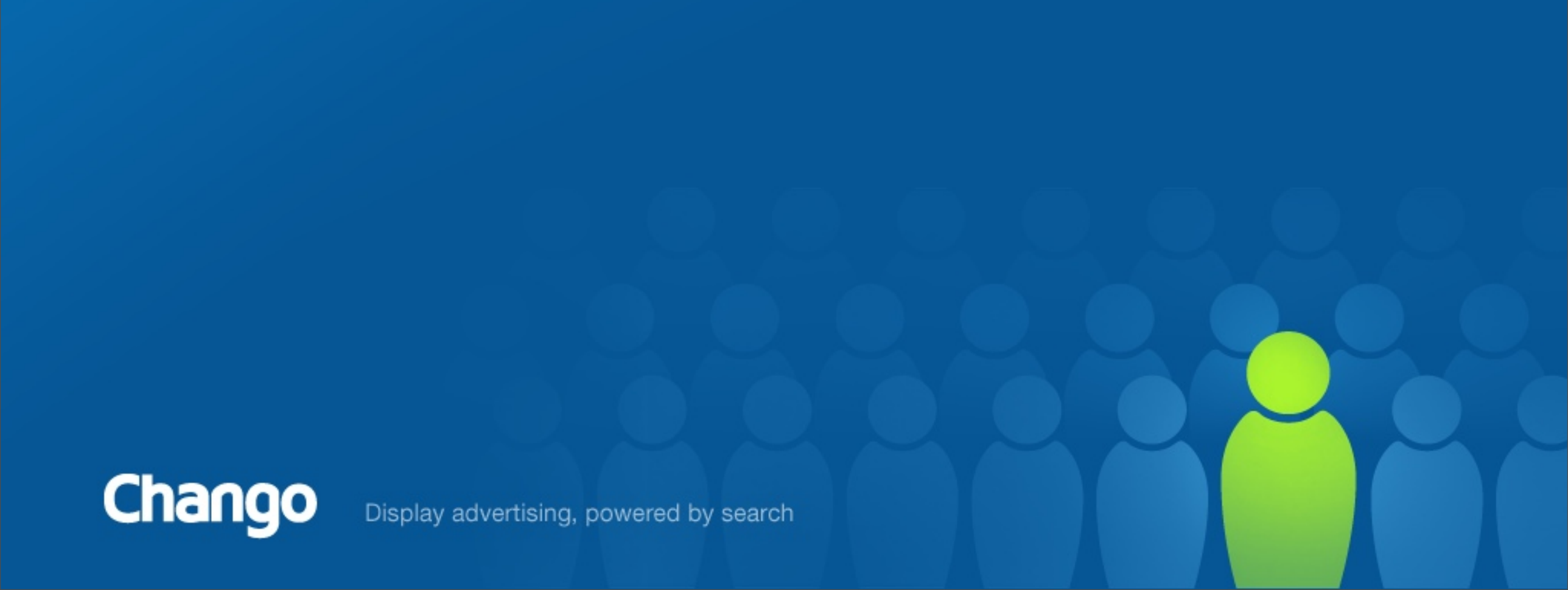
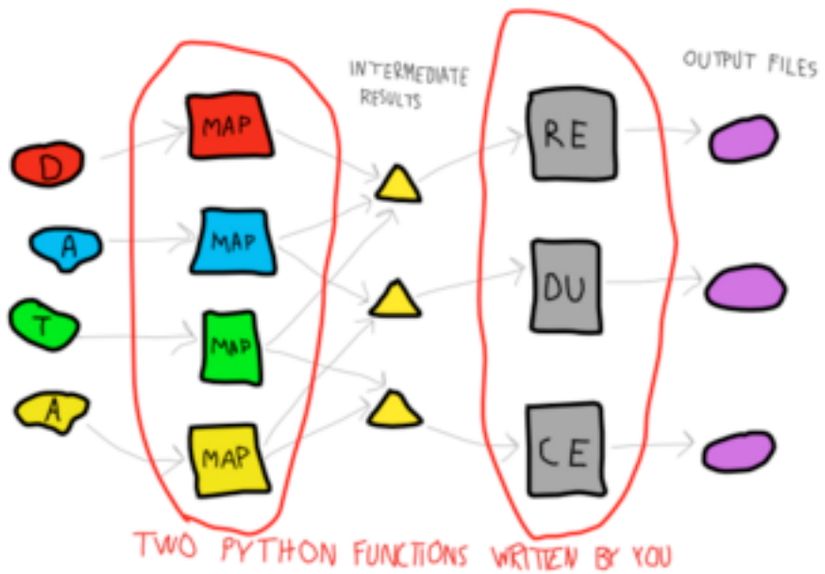
Dancing with Big Data

Inferno + Disco



Disco

- Open Source Map Reduce Platform
- 50% Erlang, 50% Python (roughly)
- Jobs are written in Python
- No Java!
- http://discoproject.com/



Why Disco?

Why Disco?

- Simplicity of Erlang Clusters
- Tag based distributed file system
- Minimal Dev-Ops Effort
- Small, readable source
- Small runtime footprint

Inferno 

- Map / Reduce Framework
  - Powered by Disco
- 100% Python (sorry)
- Developed at Chango
  - Open Sourced in March 2012

Chango

- Advertising Technology Company
  - Search Retargeting
  - Real-time bidding
- Process 10,000,000,000 records / day

Erlang at Chango

- Couchbase
  - Real-time bidding (200,000 / second)
- Disco
  - 24 Nodes (2 TB per node)

Inferno

- Query DSL for your logs
- Automation
  - E.g. Summarize to database: billions of records become1000s of rows
- Distributed computing tasks

Logs

- Structured Logs
  - Each line is valid JSON
- Replay / Reprocess Records
  - Each line has a timestamp
  - Each tag has a date
- Disco "chunks" plain text files

Example


```
{
"time":"1330969562706",
"domain":"bighealthtree.com",
"campaign_id":11056,
"search_term":"5 Signs of a Stroke You Don't
Want to Ignore",
"size":"728x90",
"ip_address":"127.0.0.1",
}
```

DEMO

Query DSL

- Rules
- Keysets
- Parts

Rules

- Automatic (Daemon Mode), Manual
- Data Source (DDFS tags)
- Date range selectors
- Processors
- Transformations

Keysets

- At least one per Rule
- Have Key and Value “Parts”
- Multiple M / R ops on the same data

Parts

- Key Parts are what you want to “map”, Value Parts are the “reduce” values
- Example: Count all the clicks for an ad on a particular site:
  - **Keys**: ad_id, site_id
  - **Values**: count (magic function)

Example

```
InfernoRule(
    map_input_stream=chunk_json_stream,
    source_tags=['adserver:chunk:clicks'],
    reduce_function=pure_maps.sorted_reduce,
    key_parts=['ad_id', 'site_id'],
    value_parts=['count'],
    field_transforms={'ad_id':to_int},
)
```

Process & Transform

- Field Transforms
- Select & Generate (Chain-able)
- Post Processors
- Input Streams (Extends Disco)

Archiving

- Update the same tag with new data
- Blobs are tagged and never reprocessed
- Tag dates are used intelligently
- Schedule data processing

DEMO

Dedication

- Jimmy Ellis, the lead singer of the hit "Disco Inferno" from '70s R&B/funk group The Trammps.
- Died March 2012 in Rock Hill, South Carolina. He was 74.

- Find us and ask questions
- http://bitbucket.org/chango/inferno
- http://inferno.rtfd.org/
- https://groups.google.com/group/python-inferno