Data : Big & Open
Big Data
Open Data

François Bancilhon
Data Publica
&
INRIA/Mobile Services Initiative
twitter.com/fbancilhon

A deluge of data

- **Lots of Data**
- **Open Data**
- **Big Data**

DATA PUBLICA

A wealth of data

- **Lots of Data**
- **Open Data**
- **Big Data**

DATA PUBLICA

Data

- Data is one of the mega trends of the 2010's
- Big Data
  - New data sets (by source and size) that require new technologies and provide new opportunities
- Open Data
  - New data sets (by source and diversity) that require new technologies and provide new opportunities

Big Data

Sources of Big Data

- Information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, posts to social media sites, cameras, microphones, digital pictures and videos posted online, RFID readers, wireless sensor networks, transaction records of online purchases, cell phone GPS signals
- 6 billion mobile phone subscriptions worldwide, and 2 billion people connected to the internet.
- The number of smartphones is increasing by 20% a year and the number of sensors by 30%.

Size of Big Data

- Terabytes (10^{12}), exabytes (10^{18}) and zettabytes (10^{21}) of data
- Everyday, we create 2.5 exabytes of data
- 90% of the data in the world today has been created in the last two years alone.
- The amount of digital information increases tenfold every five years
- Total amount of global data passed 1.2 zettabytes sometime during 2010.
- Equivalent to the amount of data that would be generated by everyone in the world posting messages on the microblogging site Twitter continuously for a century.

Size of Big Data

- When Sloan Digital Sky Survey started work in 2000, its telescope in New Mexico collected more data in its first few weeks than had been amassed in the entire history of astronomy. Now, a decade later, its archive contains 140 terabytes of information. A successor, the Large Synoptic Survey Telescope, due to come on stream in Chile in 2016, will acquire that quantity of data every five days.
- Wal-Mart, handles more than 1 million customer transactions every hour, feeding databases estimated at more than 2.5 petabytes—the equivalent of 167 times the books in America's Library of Congress
- Facebook is home to 40 billion photos.

Usage of Big Data

- Meteorology, genomics, complex physics simulations, biological and environmental research, Internet search, Web mining, finance and business informatics
- Observe, measure and predict

Usage of Big Data

- Tesco collects 1.5 billion pieces of data every month and uses them to adjust prices and promotions.
- Williams-Sonoma uses its knowledge of its 60 million customers (which includes such details as their income and the value of their houses) to produce different iterations of its catalogue.
- 30% of Amazon's sales are generated by its recommendation engine (“you may also like”).
- Placecast is developing technologies that allow them to track potential consumers and send them enticing offers when they get within a few yards of a Starbucks.

Limits of existing technology for Big Data

- Difficulties include capture, storage, search, sharing, analytics, and visualizing
- Limitations of relational databases and desktop statistics/visualization packages, requiring instead massively parallel software running on tens, hundreds, or even thousands of servers

New problems, new technology

- MapReduce
  - New framework for parallel query execution
- Hadoop
  - New architecture for parallel query execution
- NoSQL
  - New database systems

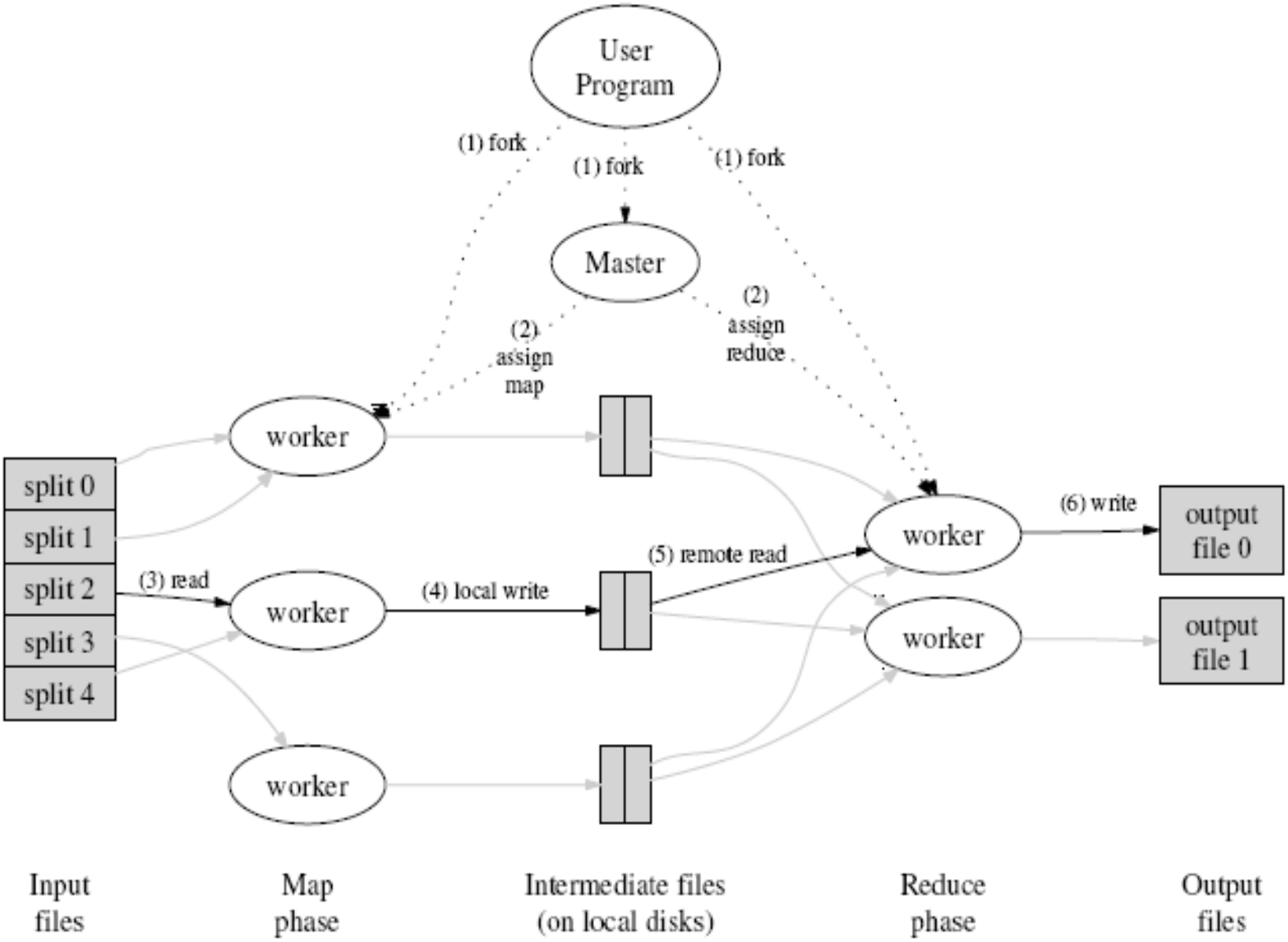
MapReduce

- Software framework introduced by Google in 2004 to support distributed computing on large data sets on clusters of computers.
- Inspired by the map and reduce functions commonly used in functional programming.
- MapReduce libraries have available in C++, C#, Erlang, Java, OCaml, Perl, Python, PHP, Ruby, F#, R, etc.

MapReduce example

```
map(String key, String value):
// key: document name
// value: document contents
for each word w in value:
  EmitIntermediate(w, "1");
reduce(String key, Iterator values):
// key: a word
// values: a list of counts
int result = 0;
for each v in values:
  result += ParseInt(v);
Emit(AsString(result));
```

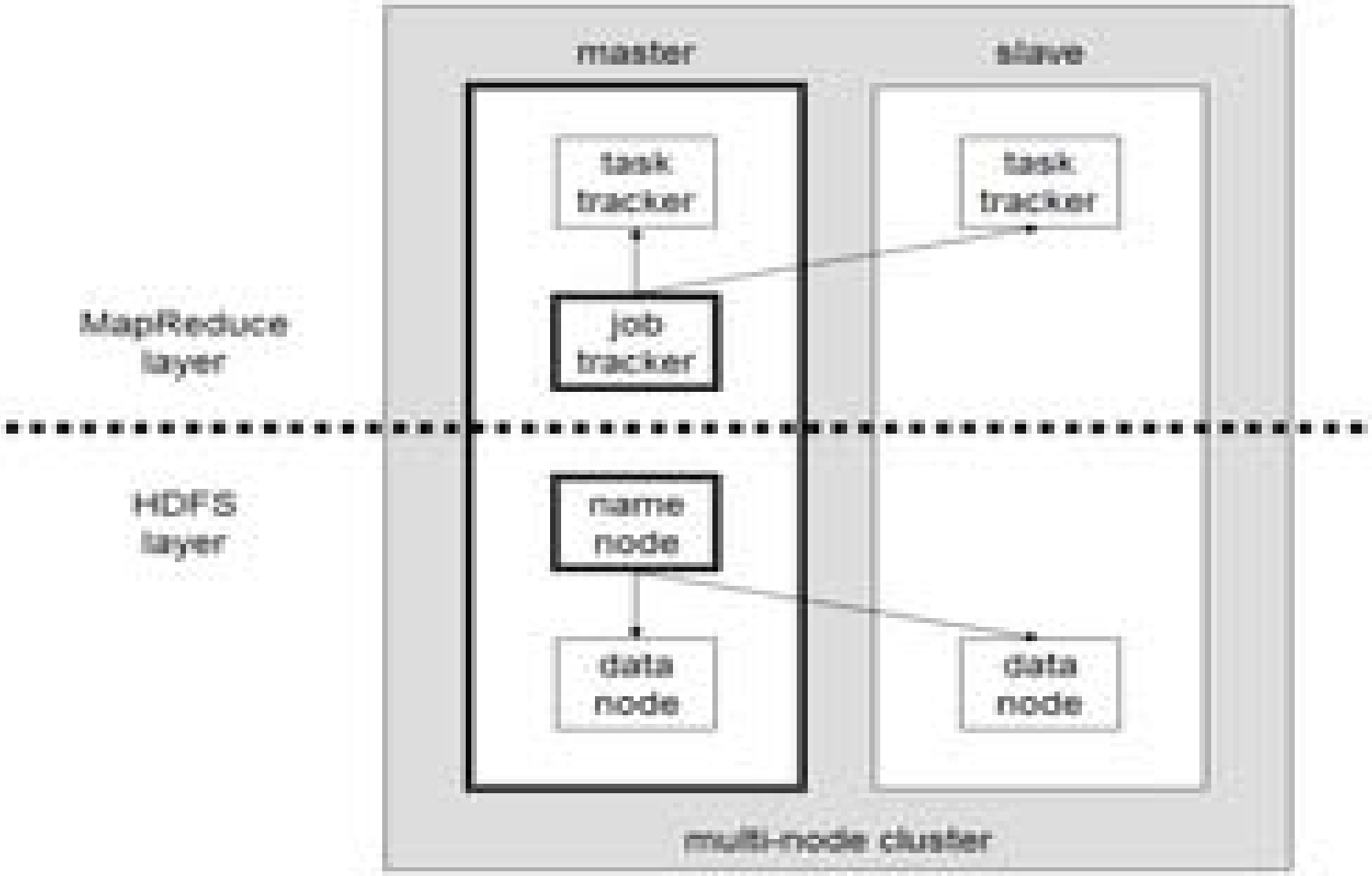
MapReduce


User Program
(1) fork
(1) fork
(1) fork
Master
(2) assign map
(2) assign reduce
worker
worker
worker
worker
worker
split 0
split 1
split 2
split 3
split 4
(3) read
(4) local write
(5) remote read
(6) write
output file 0
output file 1
Input files
Map phase
Intermediate files (on local disks)
Reduce phase
Output files

Apache Hadoop

- Software framework that supports data-intensive distributed applications under a free license.
- Enables applications to work with thousands of nodes and petabytes of data.
- Inspired by Google's MapReduce and Google File System (GFS) papers.
- Yahoo! has been the largest contributor to the project, and uses Hadoop extensively across its businesses.

Hadoop


master
slave
task tracker
task tracker
job tracker
MapReduce layer
HDFS layer
name node
data node
data node
multi-node cluster

The industrial rush on Hadoop

- Startups
  - Cloudera, Hadapt, Datastax, Mapr, Appistry, Platform Computing, etc.
- The biggies
  - IBM, EMC, Yahoo, Amazon, Google, etc.

NoSQL

- Not only SQL
- Databases (or file systems) adapted to new and big data processing
- Format (key, value) pairs
- Examples
  - Cassandra (Facebook)
    - An Apache Project
  - MongoDB (10gen)

MapReduce vs. RDBMSs

- Is MapReduce sending us back 30 years ago?
- Data independence
- Query languages (high level)
- Query optimization
- Parallel query execution

- See the debate spawned by DeWitt & Stonebraker

Open Data

Open Data

World wide move to open government data to citizens and enterprises for access and reuse

data.gov

data.gov.uk

DATA PUBLICA

Open Data: what

- Public Sector Information (PSI) made available to citizens and corporations for access and reuse
- Categories of data
  - Transportation, geography, business, sociology, ecology, environment, economy, legal, market, etc.
- Format
  - Tables, databases, xml, text, etc.
- Size
  - 300,000 table data sets in the UK ; 175,000 in France
- Restrictions
  - IP, privacy, secret

Open Data: why

- Public data, produced with taxpayer money should return to the taxpayer
- Giving public data access to citizens is one of the elements of government transparency
- PSI is an asset, government make good use of it
- PSI is fuel to the creativity of Internet and Mobile companies
- Public data can help make the life of the citizen better

Initiatives

- USA: data.gov
- UK: data.gov.uk
- Finland, Australia, Northern Ireland, New Zealand, etc
- Cities and regions: Edmonton, San Francisco, Extramadura, Toronto, etc.

Initiatives: France

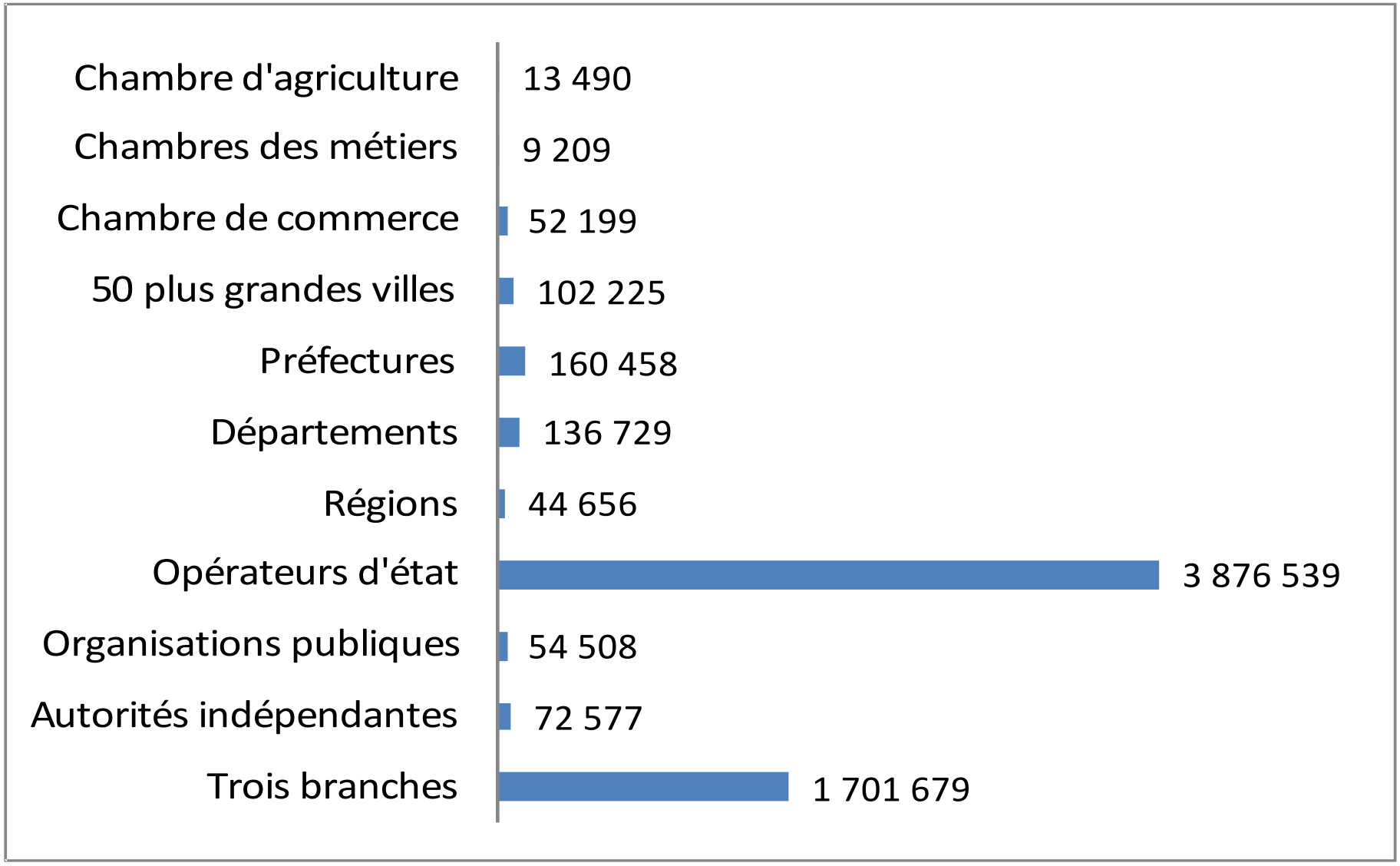
- Territories : Rennes, Paris, Bordeaux, Nantes, Marseilles
- State : APIE, Etalab (data.gouv.fr scheduled for December 2011)
- Private initiatives: Data Publica (www.data-publica.com), Regards Citoyens (www.nosdonnees.fr)

1 380 public organizations in France

- Executive, legislative, judicial
- Independent authorities
- Public organizaitions
- State operators
- Regions, departments, urban communities
- 50 largest cities
- Chamber of commerce, agriculture and craftsmen

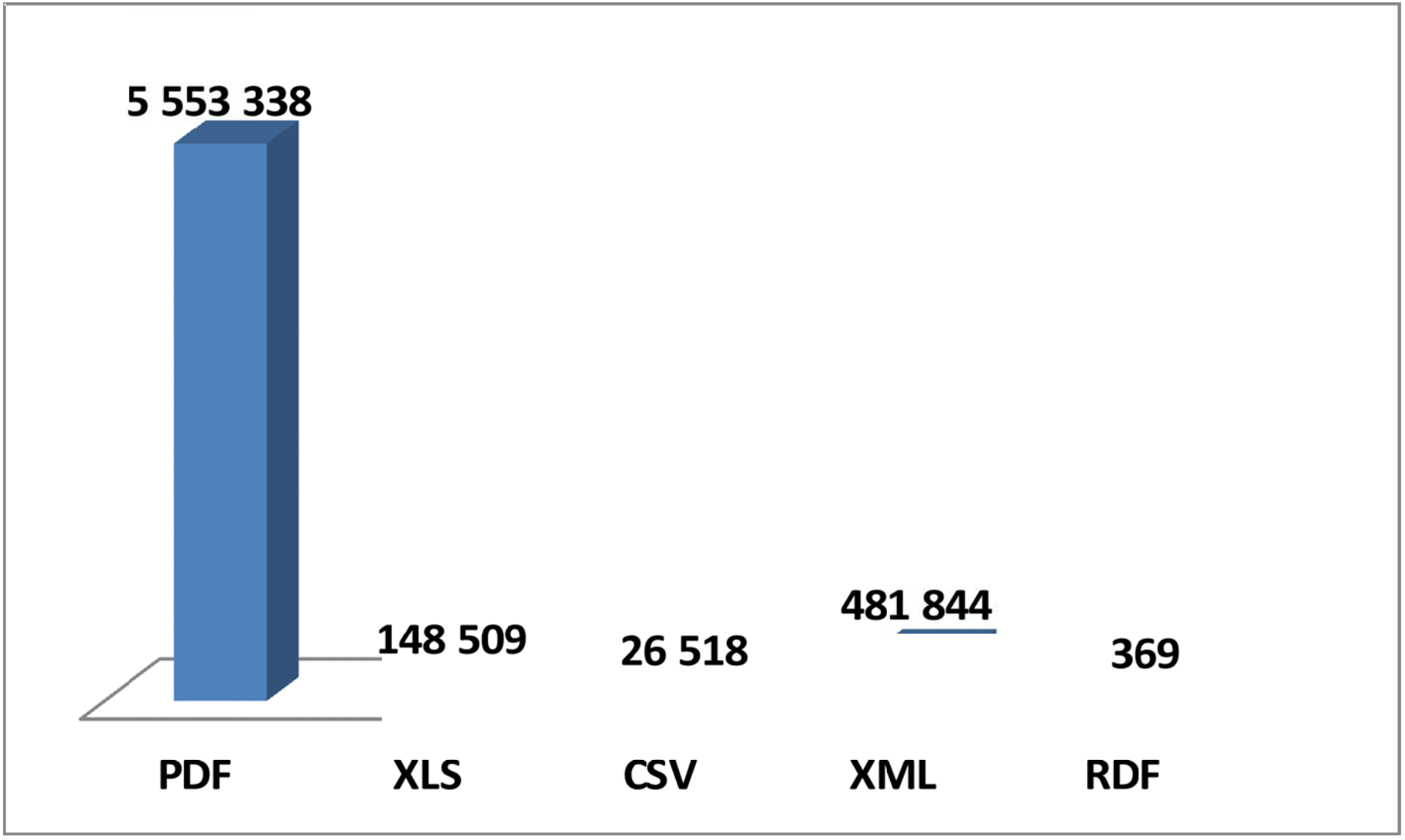
DATA PUBLICA

How much public data: 6.5M files



DATA PUBLICA

In which formats?



DATA PUBLICA

On what topics?

DATA PUBLICA

The Data geek

- A new kind of professional has emerged, the data scientist
  - combines the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data.
  - Job of statistician will become the “sexiest” around
  - Data Journalist « the press will have to become data savvy »

Conclusions

- Data is a major trend
- Big Data and Open Data
  - Technology, business, society, politics
  - A major opportunity