Experience with Log-data Analysis and the Use of Big Data Technologies in Large-Scale Assessment in France

Saskia Keskpaik

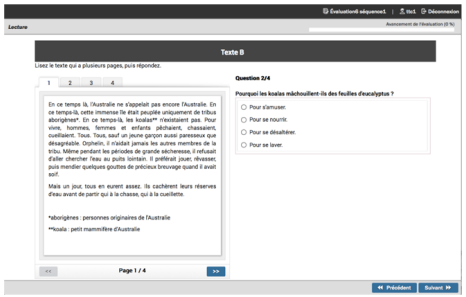
DEPP, Ministry of Education, France

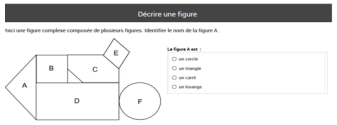
DEPP AND ASSESSMENTS

- DEPP is responsible for 3 types of assessments:
  - National census-based assessments
  - National sample-based assessments
  - International assessments (ICILS, PIRLS, PISA, TIMSS)
- Produces indicators for classroom, local/regional and national use levels
- Traditionally paper-based, transition to digital assessment

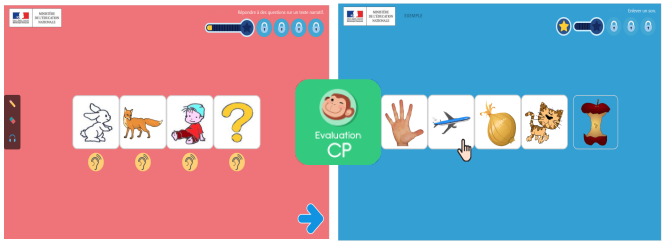
DIGITAL TRANSITION

- Middle and high schools:
  - Online platform TAO on desktop computers





- Primary schools:
  - Offline app on tablets



DIGITAL ASSESSMENTS

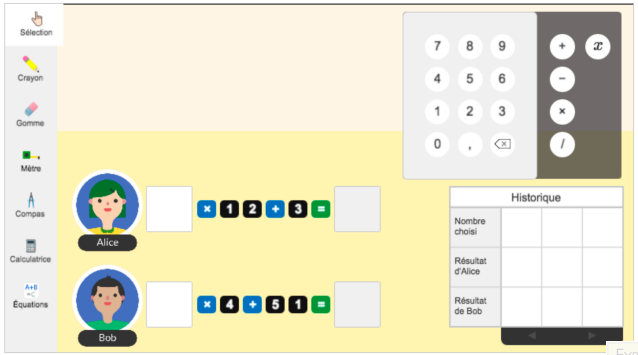
- Opportunities/challenges:
  - New opportunities: cost reduction, multimedia, accessibility, interactivity, simulated situations, automatic coding, adaptive testing, process-data, etc.
  - New challenges: comparability with pencil&paper (trends), usability, security, confidentiality, data storage, data processing, etc.

- 2 major aspects:
  - Scalability
  - Innovative assessments

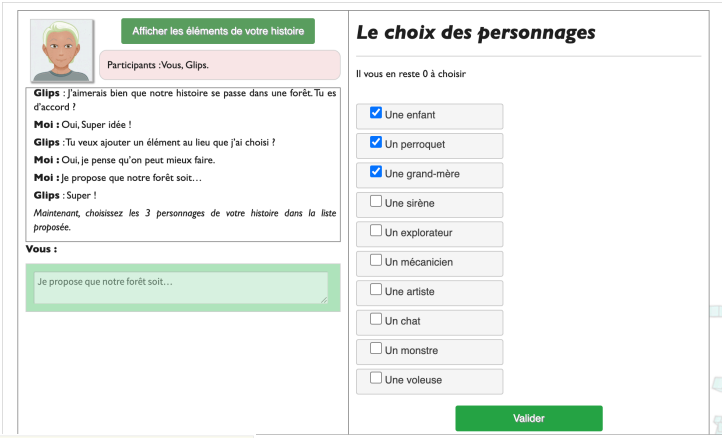
INNOVATIVE ASSESSMENTS

Examples of interactive items

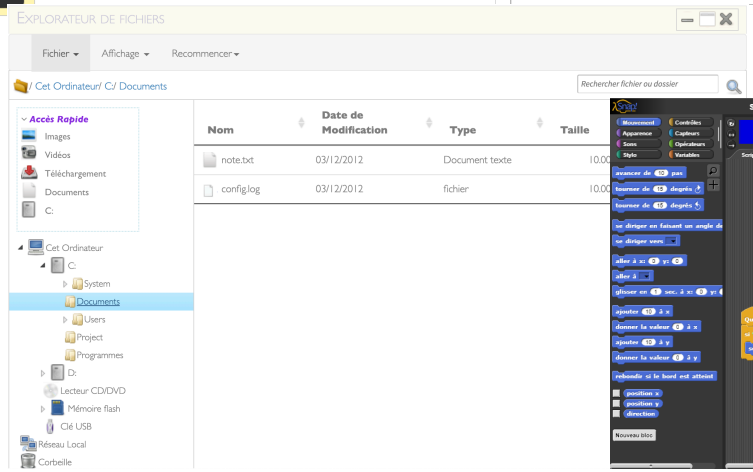
Maths



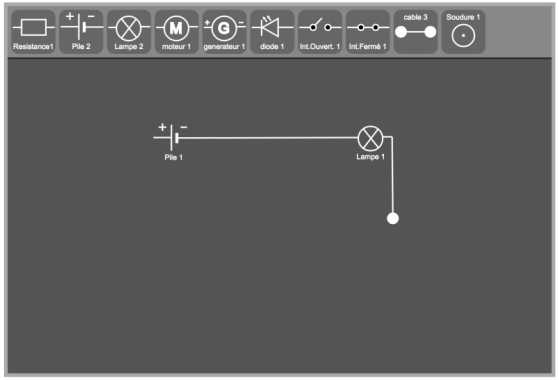
ChatBot



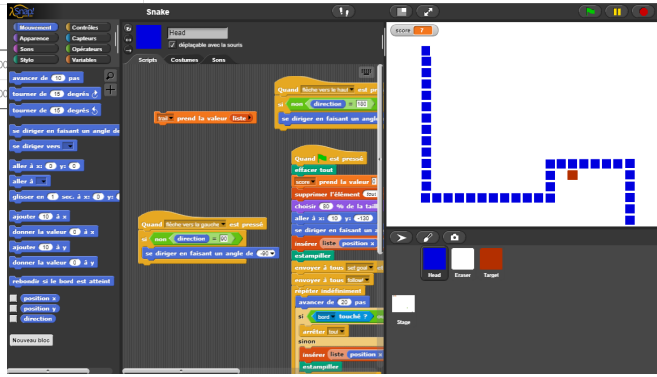
File Explorer



Physics



Coding



GROWING DATA

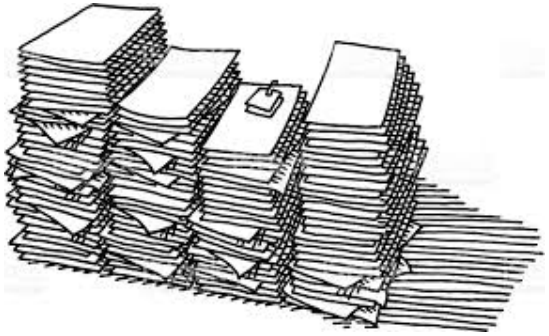
- Interactive items generate very large data
- Number of test-takers and interactive items expected to grow rapidly
- Experimentation of interactive items in 2017: 9 GB of data for a total of 15,000 students (3000 students per item)

➔Need to look for "Big Data" solutions

WHAT IS BIG DATA ?

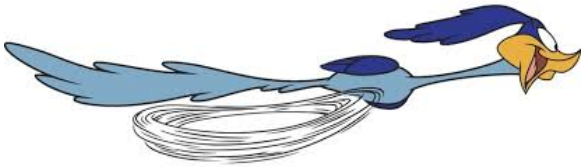
- The 3 **V**s

Volume



Variety

Velocity



- Collection and storage of data so large and complex that it exceeds the processing capabilities of existing techniques and systems

VOLUME

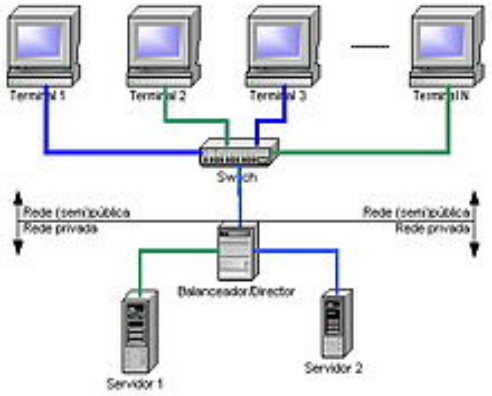
- Terabyte (1000 GB) or even petabyte (1,000,000 GB) data
- In our case small but growing Big Data:
  - 17PCIX : 15,000 test-takers, 9 GB
  - EVA6 : 800,000 test-takers, 30 GB
  - ELAINE (DEPP B4) : between 5 and 13 TB (experimental phase)
- Problematic processing on a local machine for 2 GB files and over (with R software)

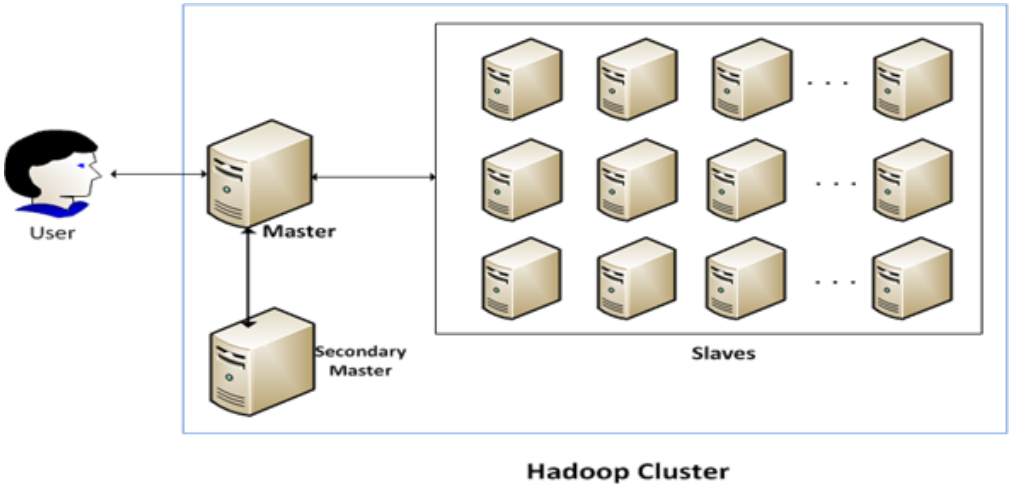
VARIETY

- Unstructured (text, images, sounds…) or partially structured (JSON, XML…) data.
- In our case mostly semi-structured but complexity increasing:
  - Process data: JSON, “pseudo-json”
  - Fluency assessment: sounds
- Data with “noise”

DISTRIBUTED INFRASTRUCTURE

- Data distribution and data processing on several machines: the "cluster"
- Master node and Worker nodes

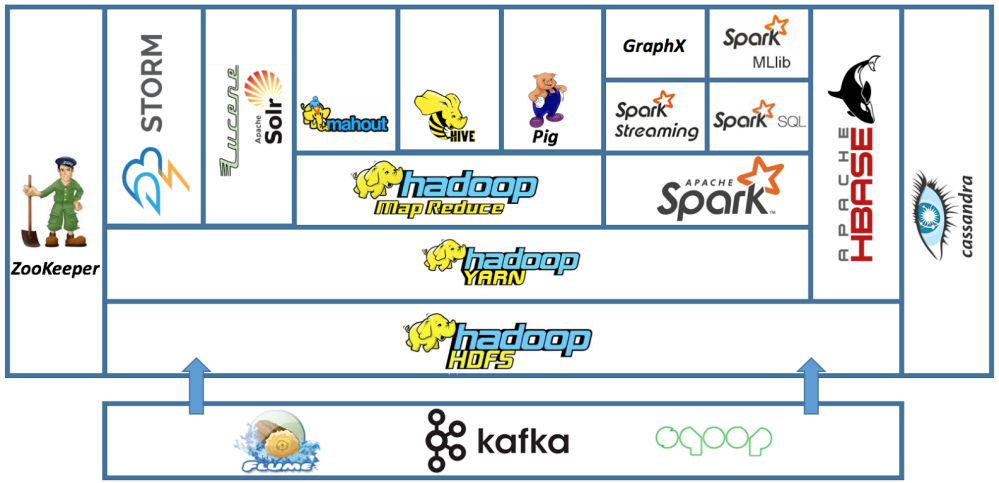




Hadoop Cluster

BIG DATA ARCHITECTURE

- Hadoop & Spark frameworks
- **Hadoop** – free and open source framework, designed to deal with huge volumes of data, in a distributed environment.



HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

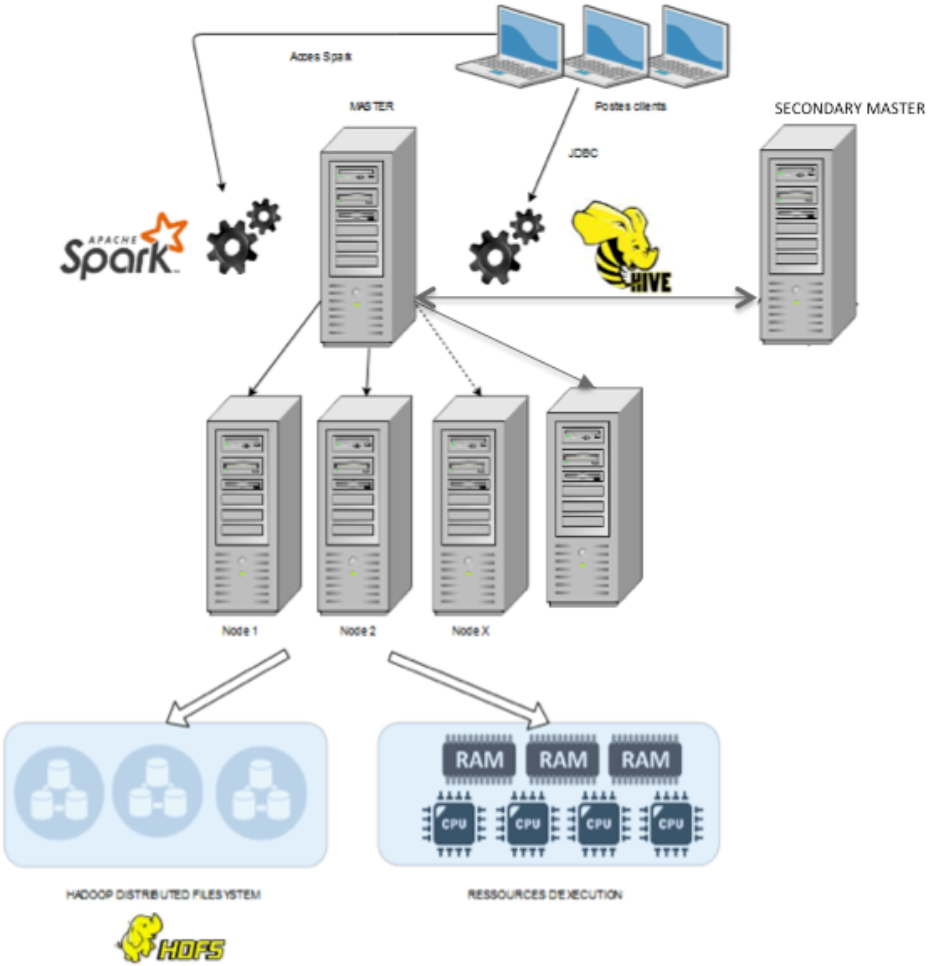
- Distributed, scalable and portable file system, written in java for the Hadoop framework
- Segmentation of data into blocks:
  - Block size is set to 128 MB by default, but configurable
  - 10 GB file = 10*1024/128 = 80 blocks
- Replication and distribution of data on several cluster nodes (default by 3)

  ➔ ensures data integrity even if a node fails
- Optimized for large files

SPARK

- Spark is an open source distributed computing framework
- Does not have its own file management system - uses HDFS
- Basic concept: RDD (resilient distributed dataset)
- Calculations within RAM: speed
- User interface in Python, R : PySpark/SparkR (+ sparklyR)
- Several libraries + use of Python/R libraries

DEPP'S BIG DATA ARCHITECTURE



GENERAL DATA PRE-PROCESSING

- Extract the logs from the raw data: csv file

  => moving towards data base (clone) queries

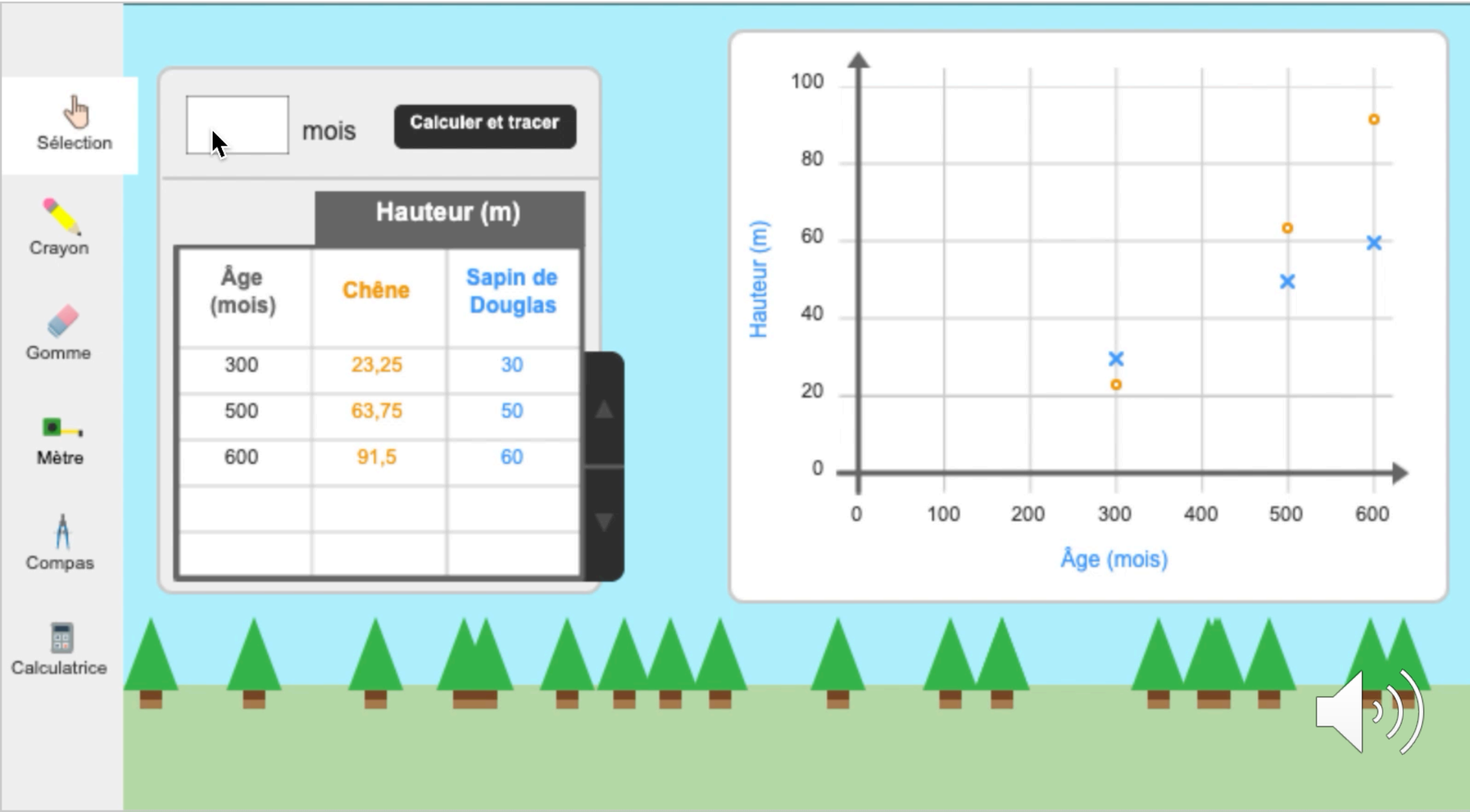
- Decode the data: double encoding in base64 and lzstring

  https://github.com/pieroxy/lz-string

- PySpark and parallel processing for all items

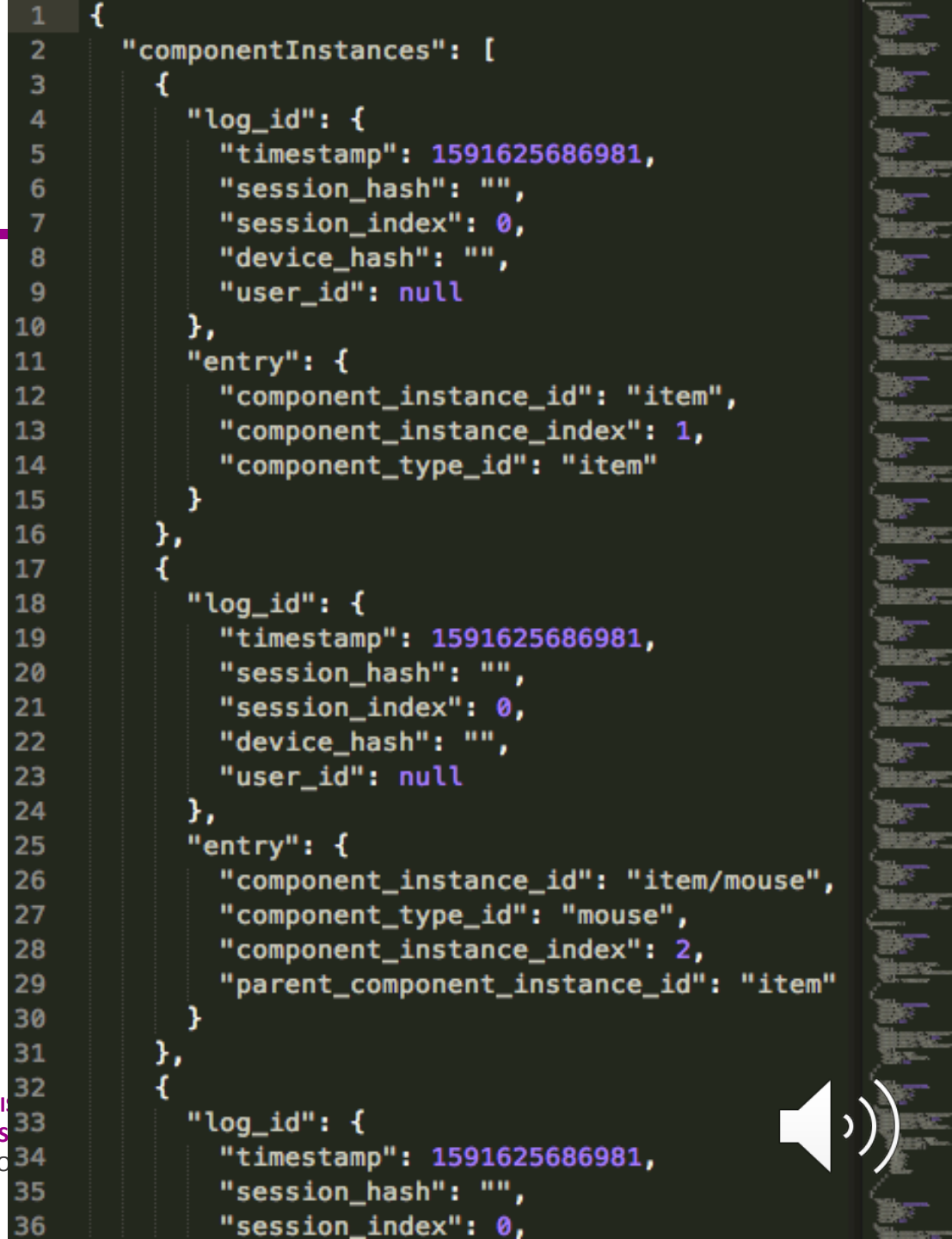
- Store data on HDFS

ITEM-SPECIFIC DATA PROCESSING



LOG-FILE STRUCTURE

- 30 sec of interaction = 4746 lines of json data
- Events and State of each logged element
- Timestamp of events and overall start and end timestamp

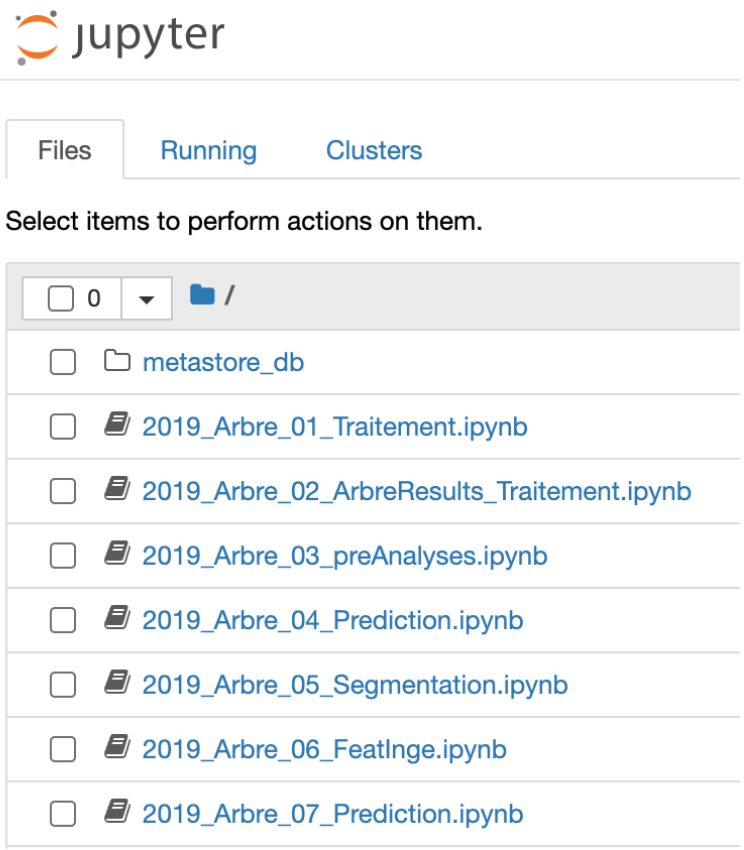


ITEM-SPECIFIC WORKFLOW (1)

- Notebooks for each stage (data preparation, pre-analysis, ML)
- Start out with json data (Spark SQL module)

```
dfRaw.printSchema()
```

```
root
 |-- contents: struct (nullable = true)
 |     |-- componentEvents: array (nullable = true)
 |     |     |-- element: struct (containsNull = true)
 |     |     |     |-- entry: struct (nullable = true)
 |     |     |     |     |-- component_event_id: string (nullable = true)
 |     |     |     |     |-- component_instance_id: string (nullable = true)
 |     |     |     |     |-- component_instance_index: string (nullable = true)
 |     |     |     |     |-- component_type_id: string (nullable = true)
 |     |     |     |     |-- data: struct (nullable = true)
 |     |     |     |     |     |-- center_x: long (nullable = true)
 |     |     |     |     |     |-- center_y: long (nullable = true)
 |     |     |     |     |     |-- char: string (nullable = true)
 |     |     |     |     |     |-- col: long (nullable = true)
 |     |     |     |     |     |-- currentStep: long (nullable = true)
 |     |     |     |     |     |-- direction: string (nullable = true)
 |     |     |     |     |     |-- draw_el_x: long (nullable = true)
 |     |     |     |     |     |-- draw_el_y: long (nullable = true)
 |     |     |     |     |     |-- draw_id: long (nullable = true)
```



ITEM-SPECIFIC WORKFLOW (2)

- Gradual unwrapping of variables of interest

```
exploded.registerTempTable('exploded')
exploded.select('json_id','student_id','timestamp','comp_type','comp_event','position_data','x','value').show()
```

```
+--------+----------+-------------+-------------+--------------------+--------------------+----+-----+
| json_id|student_id|    timestamp|    comp_type|          comp_event|       position_data|   x|value|
+--------+----------+-------------+-------------+--------------------+--------------------+----+-----+
|00104611|  0510029E|1494327816139|        mouse| MouseMoveStartEvent|                null| 270| null|
|00104611|  0510029E|1494327816377|        mouse|   MouseMoveEndEvent|  M271,5L271,5L271,5| 271| null|
|00104611|  0510029E|1494327817276|        mouse| MouseMoveStartEvent|                null| 274| null|
|00104611|  0510029E|1494327817608|        mouse|   MouseMoveEndEvent|M430,85L473,69L47...| 473| null|
|00104611|  0510029E|1494327817610|        mouse| MouseMoveStartEvent|                null| 474| null|
|00104611|  0510029E|1494327818891|        mouse|   MouseMoveEndEvent|M532,86L573,101L6...| 696| null|
|00104611|  0510029E|1494327819257|        mouse| MouseMoveStartEvent|                null| 693| null|
|00104611|  0510029E|1494327820224|        mouse|   MouseMoveEndEvent|M674,176L654,193L...|  40| null|
|00104611|  0510029E|1494327820292|        mouse| MouseMoveStartEvent|                null|  35| null|
|00104611|  0510029E|1494327821160|        mouse|   MouseMoveEndEvent|M63,380L83,299L12...| 120| null|
|00104611|  0510029E|1494327821176|        mouse|      MouseDownEvent|                null| 120| null|
|00104611|  0510029E|1494327821176|        mouse|MouseMoveContinue...|                    | 120| null|
|00104611|  0510029E|1494327821176|        mouse|MouseMoveContinue...|                null| 120| null|
|00104611|  0510029E|1494327821177|graphingTable|ArbreGraphDeselec...|                null|null| null|
|00104611|  0510029E|1494327821178|treeDataTable|ArbreTableDeselec...|                null|null| null|
|00104611|  0510029E|1494327821275|        mouse|        MouseUpEvent|                null| 120| null|
|00104611|  0510029E|1494327821275|        mouse|MouseMoveContinue...|                    | 120| null|
|00104611|  0510029E|1494327821275|        mouse|MouseMoveContinue...|                null| 120| null|
|00104611|  0510029E|1494327821275|  numberInput|       inputSelected|                null|null| null|
|00104611|  0510029E|1494327821276|    undefined|          SHOW_EVENT|                null|null| null|
+--------+----------+-------------+-------------+--------------------+--------------------+----+-----+
only showing top 20 rows
```

ITEM-SPECIFIC WORKFLOW (3)

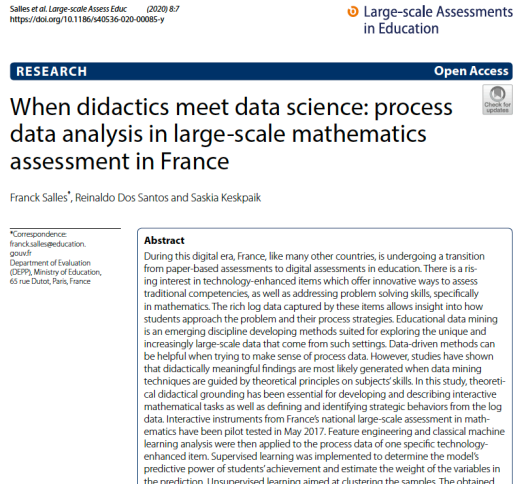
- Feature engineering to create new meaningful variables
- Output: tabular data with useful variables for analysis
- Next step: analysis with python pandas, numpy, scikit-learn

```
: df_treeData.show()
```

json_id	student_id	months_list	months_list_length	sequence	nb_alternance	contain_g	nb_apres_g
00270644	0142T7E	[700, 450, 300, 2...	10	mmmmpppppg	2	1	0
00007014	0350882U	[400, 395, 390]	3	mmg	1	1	0
00887273	03311E7L	[200, 150, 125, 1...	13	mmmmmppmmmmmg	3	1	0
01032603	0831K19N	[100, 50, 200, 25...	10	mppppmpmpg	6	1	0
00529779	0789Q23H	[350, 360, 370, 3...	5	ppppg	1	1	0
00590188	0782Z3B	[10, 100, 400, 35...	7	ppmpppg	3	1	0
00729375	0941S20L	[700]	1		0	0	-1
00001541	0330083L	[100, 390, 400]	3	pgp	1	1	1
00597459	0443P9X	[7, 720, 5, 2, 1,...	7	pmmmpm	4	0	-1
00596166	0841W19K	[500, 700, 700, 9...	14	pepmmmmpppppp	3	0	-1
00905274	0321M7E	[360, 370, 380, 390]	4	pppg	1	1	0
00469067	0336K12L	[300, 400, 450, 3...	11	ppmmmmpmpp	5	0	-1
00598597	0443P14R	[400, 399, 401, 4...	5	mppm	3	0	-1
00625193	0852P7T	[8, 50, 250, 350,...	11	pppppmpmmp	5	0	-1
00535898	0521Z20Q	[300, 390]	2	pg	1	1	0
01079264	0720068H	[350, 368, 360, 3...	12	pmppmmmmmmmg	4	1	0
00052968	0684N19U	[8, 9, 12, 200, 3...	9	ppppppppg	1	1	0
00789138	0374S6Y	[400, 401, 400]	3	pm	2	0	-1
00277866	0931485J	[800, 390]	2	mg	1	1	0
00186120	0375U5A	[400, 399, 401, 3...	5	mpmmg	3	1	0

only showing top 20 rows

FURTHER ANALYSIS


Salles et al. Large-scale Assess Educ (2020) 8:7
https://doi.org/10.1186/s40536-020-00085-y

Large-scale Assessments in Education

RESEARCH Open Access

When didactics meet data science: process data analysis in large-scale mathematics assessment in France

Franck Salles*, Reinaldo Dos Santos and Saskia Keskpaik

*Correspondence:
franck.salles@education.gouv.fr
Department of Evaluation (DEPP), Ministry of Education, 65 rue Dutot, Paris, France

Abstract

During this digital era, France, like many other countries, is undergoing a transition from paper-based assessments to digital assessments in education. There is a rising interest in technology-enhanced items which offer innovative ways to assess traditional competencies, as well as addressing problem solving skills, specifically in mathematics. The rich log data captured by these items allows insight into how students approach the problem and their process strategies. Educational data mining is an emerging discipline developing methods suited for exploring the unique and increasingly large-scale data that come from such settings. Data-driven methods can be helpful when trying to make sense of process data. However, studies have shown that didactically meaningful findings are most likely generated when data mining techniques are guided by theoretical principles on subjects' skills. In this study, theoretical didactical grounding has been essential for developing and describing interactive mathematical tasks as well as defining and identifying strategic behaviors from the log data. Interactive instruments from France's national large-scale assessment in mathematics have been pilot tested in May 2017. Feature engineering and classical machine learning analysis were then applied to the process data of one specific technology-enhanced item. Supervised learning was implemented to determine the model's predictive power of students' achievement and estimate the weight of the variables in the prediction. Unsupervised learning aimed at clustering the samples. The obtained

Salles, F., Dos Santos, R. Keskpaik, S.

When didactics meet data science: process data analysis in large-scale mathematics assessment in France

https://rdcu.be/b4vqf

CHALLENGES AND FUTURE PROJECTS

- Decisions taken from item development point of view might not be optimal from data processing point of view: strengthen collaboration at item development level
- Multitude of log structures: standardization
- Industrialization of data processing (collaboration with CITO, distributed ML algorithms: Spark MLlib)
- Integration to other workflow stages: towards an assessment ecosystem

THANK YOU!

saskia.keskpaik@education.gouv.fr