Telecom Valley

Soirée du <Test Logiciel>

Big data testing

Debdatta DEY

All4Test

09/12/2020

Introduction

Debdatta DEY

Test Automation - Big Data consultant

ALL4TEST Team

Working on Kering IT project

ALL4TEST is a french pure player in QA & testing

Agenda:

- What is Big Data Testing and why do we need it?
- How to adapt big data in test?
- An overview of daily big data processing in the field of CRM
- The benefits of big data testing
- The needs of the test environment
- Real-time data ingestion and flow in the use of Nifi & Kafka in a real CRM
- A scenario for validating test data using user acceptance criteria
- The challenges of Big Data Testing in our project
- How do we try to overcome them?
- GDPR and Big Data?
- Provide realistic data for our software testing

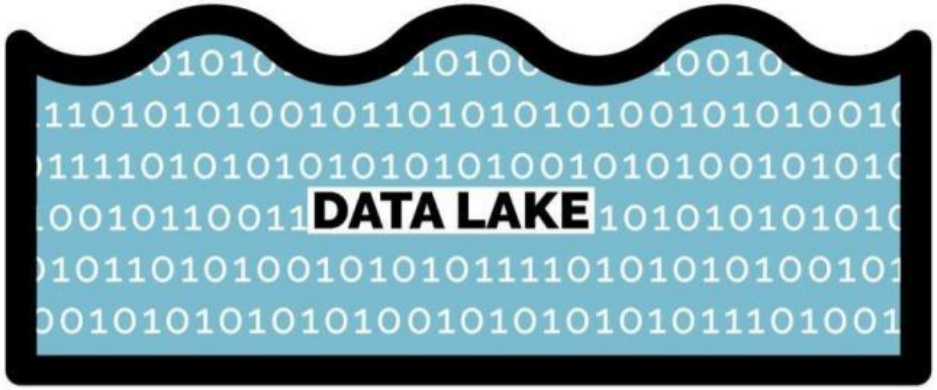
Before jumping to the details

Five words for big data...

Data analytics

Data Analytics is a science of examining raw data, with the aim of drawing conclusions from that information.

Data lake

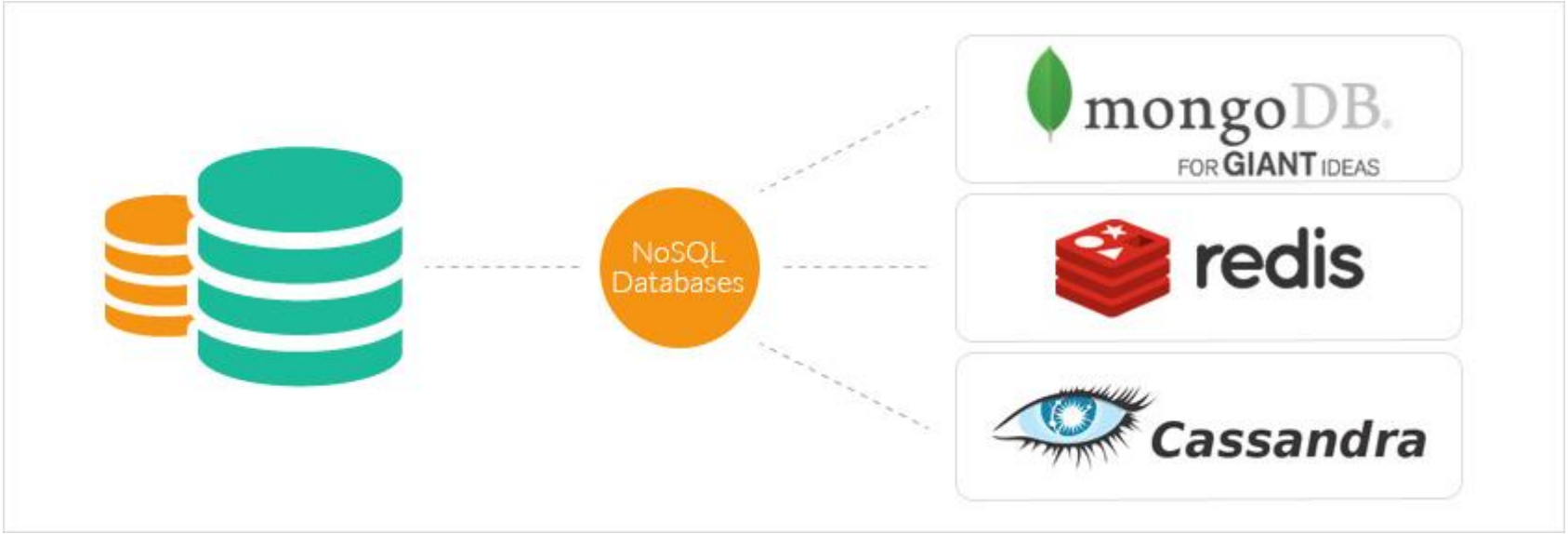


A **data lake is** a centralized storage location that contains big data in a raw and granular format from a large number of sources.

Data scientist

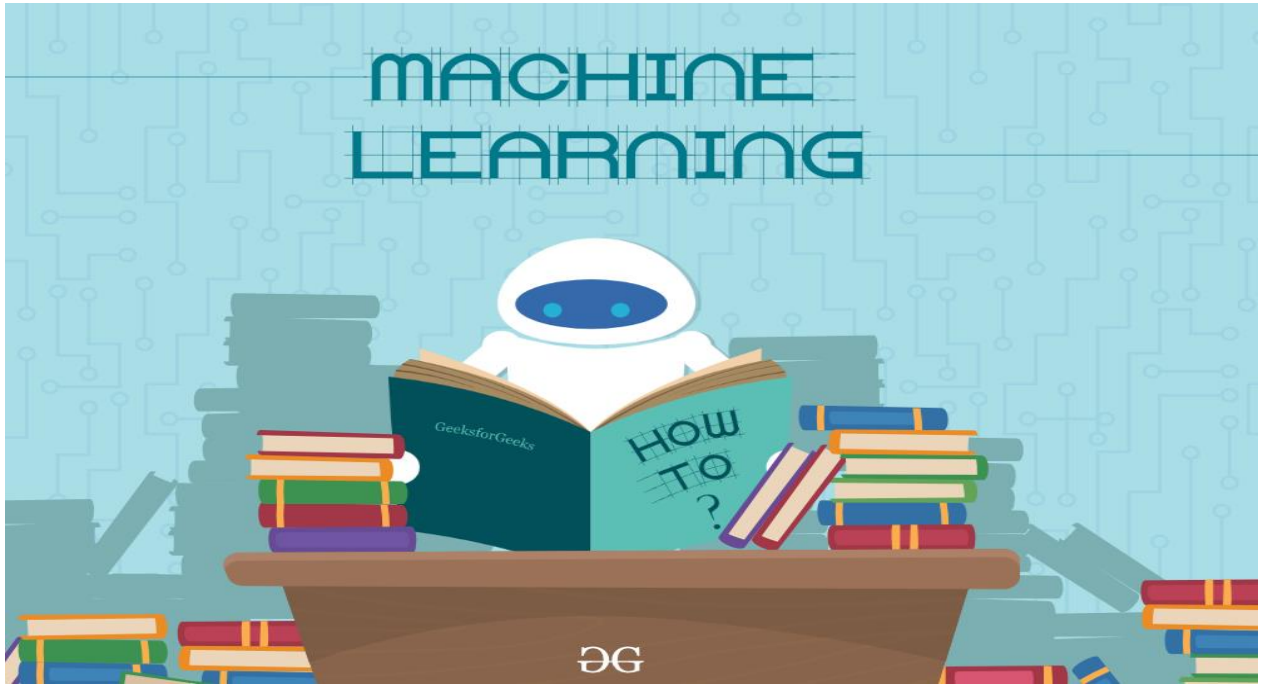
The **Data Scientist** is an expert in the management and in-depth analysis of massive data.

NoSQL



Not only SQL

Machine learning



Instead of programming the computer every step of the way, this approach gives the computer instructions that allow it to learn from data without new step-by-step instructions by the programmer.

What is big data & why do we use it?

"*One of the most important uses of big data is to generate small data .*
Small data refers to data that do not have volume, velocity, and variety issues."

- Cost saving
- Time reducing
- Efficiency of the business
- More accuracy

"Consumer data will be the biggest differentiator in the next two or three years. Whoever unlocks the reams of data and uses it strategically will win".

Areas of big data:

What is big data testing

The key factor for Big data testing

- Focus on verification of its data processing.
- Performance and functional testing.
- High level of testing skills.
- Data validation.
- Business logic validation.
- Output validation.

Data validation:

- Correct data is pulled
- Correct data is extracted
- Compare source data to the final data

Process validation:

- Business process validation
- Data accuracy
- Key pairs are generate
- Map reduce process works correctly

Output validation:

- Check the transformation rules are correctly applied
- Check the data integrity and successful data load into the target system
- Check that there is no data corruption by comparing the target data with the destination data.

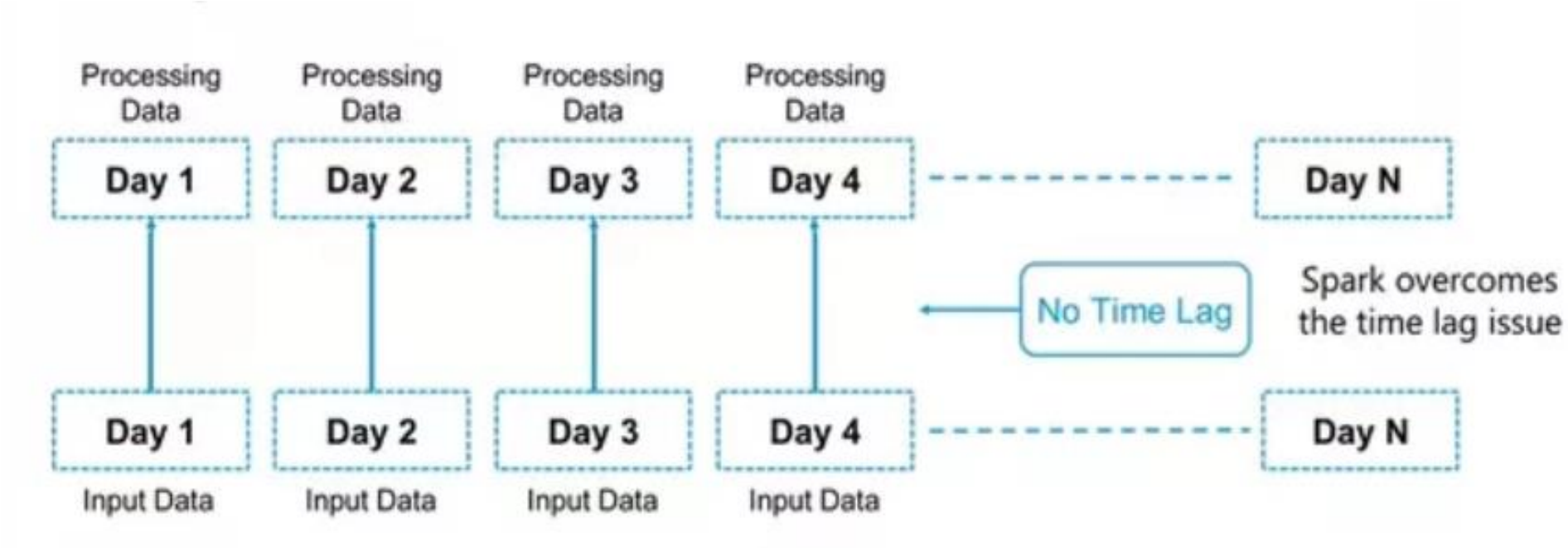
An overview of big data processing in daily CRM fields

Testing Big Data application is more verification of its data processing rather than testing the individual features of the software product.

- Real Time processing.
- Batch processing.

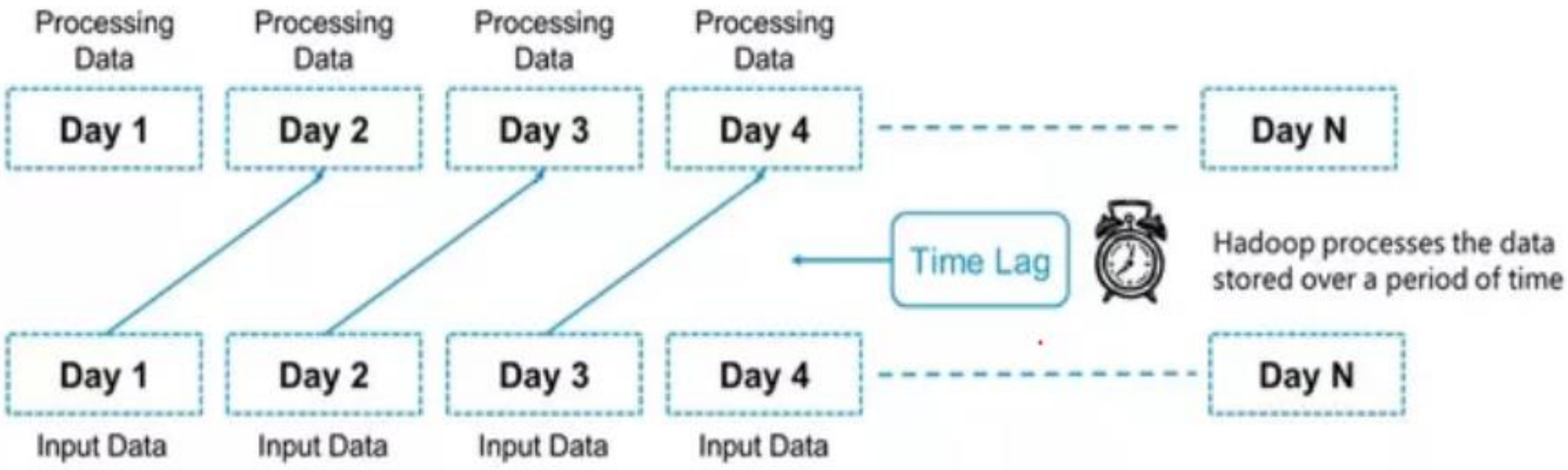
Real Time Processing

Real time data processing involves a continual input, process and output of data. Data must be processed in a small time period or near real time.



Batch processing

Where the data processing happens of blocks of data where that have already have been stored over a period of time.



The benefits of big data testing

Big data testing helps you find the qualitative, accurate and intact data.

- Data Accuracy
- Cost-effective Storage Effective Decision-making And Business Strategy
- Right Data At The Right Time
- Reduces Deficit and Boosts Profits

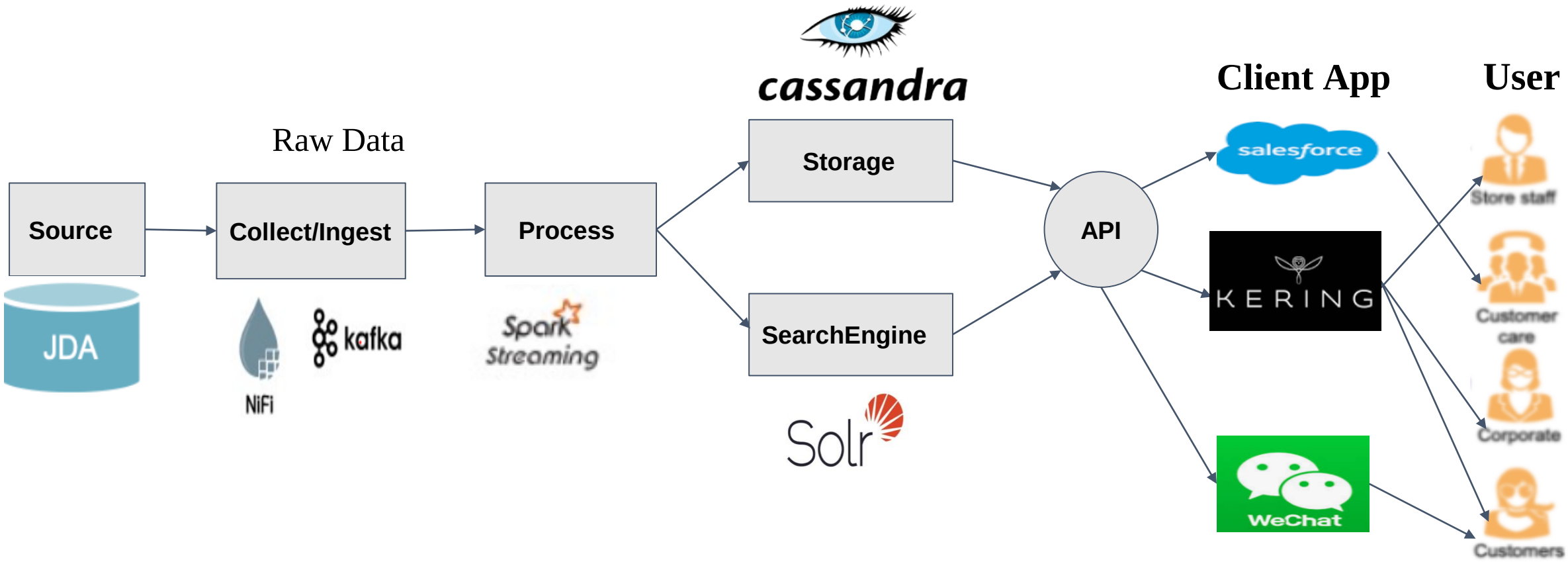
Test Environment Needs:

Test Environment needs to depend on the type of application you are testing.

For Big data testing, the test environment :

- It should have enough space for storage and process a large amount of data
- It should have a cluster with distributed nodes and data
- It should have minimum CPU and memory utilization to keep performance high.

Data ingestion and flow in CRM



Data ingestion:

- We take the real data put it into parking for few days.
- We run the batch to store the data.
- We use nifi or kafka to listen to the real time .
- Run the Spark job for batch & real time data ingestion into cassandra and solr database.
- Data is distributed according to row key hash.
- Data search through API in json format.
- Verify the source and destination data are same.

Cassandra(High volume of data across distributed system) gives fast read while solr is full-text based search

A test scenario of data validation using user acceptance criteria:

I want to search for customer data and retrieve from source so that I can use it in my application for real time and batch.

Given we check my application doesn’t have that specific customer as prerequisite.
When we read the data in KAFKA(A middleware message carrier) sent by enterprise source in real time.
And Run the spark job we absorb the data in cassandra and solr.
Then check the data is injected properly in cassandra and solr.
And also compare the data in source and destination to have the data integrity.

Challenges in Big Data Testing In our project:

- Big volume of data and Heterogeneity.
- Data sanity.
- Unstable environment.
- Lack of technical expertise.
- Effective communication & collaboration teams for test automation.
- Selecting right tools.
- Selecting proper test approach.
- Stretched deadline and costs.

How we are trying to overcome:

Big Data processing is a very promising field in today's complex business environment.

- Applying the right dose of test strategies.
- Following best practices would help ensure qualitative software testing.
- Recognise and identify the defects in the early stages of testing and rectify them.
- Testing approaches are all driven by data.

Did GDPR take a bite out of big data?

General Data Protection Regulation (GDPR)—perhaps the most famous data privacy regulation ever—came into effect on May 25, 2018.

For a quick refresher: GDPR rewrote the rulebook for how businesses can use EU consumer data for things like analytics. GDPR has shunted responsibility for data privacy away from individual consumers and firmly to the business, bringing global attention to how businesses handle the processing of personally identifiable information (PII).

Regardless of size, 93 percent of small businesses say they have prepared in some way to comply using ***Anonymization*** .
Anonymization is the ability for the data controller to anonymize the data in a way that it is important.

Providing realistic data for our software testing:

- Accelerating the software testing process is key to improving.
- Automated tools can test the software,
- Anonymize the data, this real data is safe to process.
- Today customers are all called ***Anonymized Anonymized***.
- Shuffling of the letters for Name, email address and phone number.

As enterprise source has a technical constraint, it will be more easier for the physical address to keep only the country.

But it can be a problem for some team to have only country, and no address line, not zip, no city .

We are looking for an effective solution for that

Thank you

Thank you

Thank you

Thank you for your attention.

Q & A?

Thanks you ! Any Question ?

Contact us on : contact@all4test.com

French version of the presentation :

https://www.all4test.fr/dossiers-thematiques/comment-tester-les-systemes-big-data-data-lake/