Grid Dynamics

Алексей Лянгузов
Grid Dynamics, SQA Architect

Контроль качества в BigData проектах

SQA DAYS#23

План

О чем будет данная презентация

- Что такое Big Data?
- Контроль качества в Big Data системах
  - Качество Данных
  - Функциональное Качество
  - Качество ”Стендов”
  - Нефункциональное Качество
  - Инструментарий
- Что нужно тестировщику, чтобы работать с Big Data

Что такое Big Data

Введение

Что же такое Big Data

Big Data – это …

Введение

Что же такое Big Data

Big Data – это … я вам не скажу, что это такое!

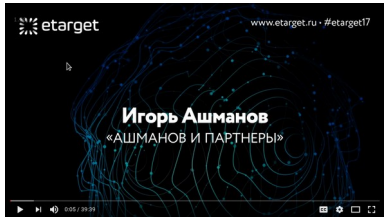
Введение

Что же такое Big Data

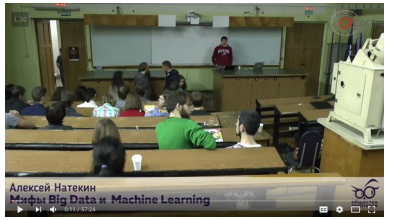
Big Data – это … я вам не скажу, что это такое!

А вот они расскажут

Владимир Красильщик
Анти-введение в Big Data



Игорь Ашманов
Аналитика Big Data

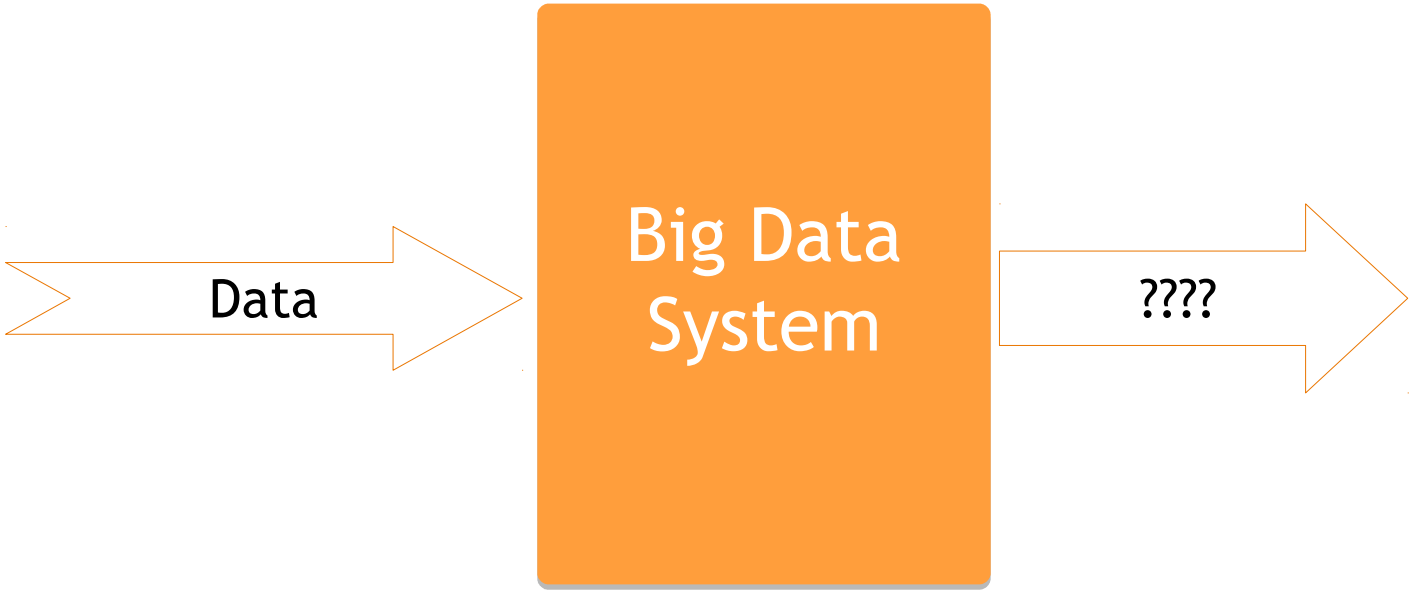


Алексей Натекин
Мифы Big Data и ML

А мы с вами лучше поговорим о том, что такое Big Data System
или
Система, работающая с Большими Данными

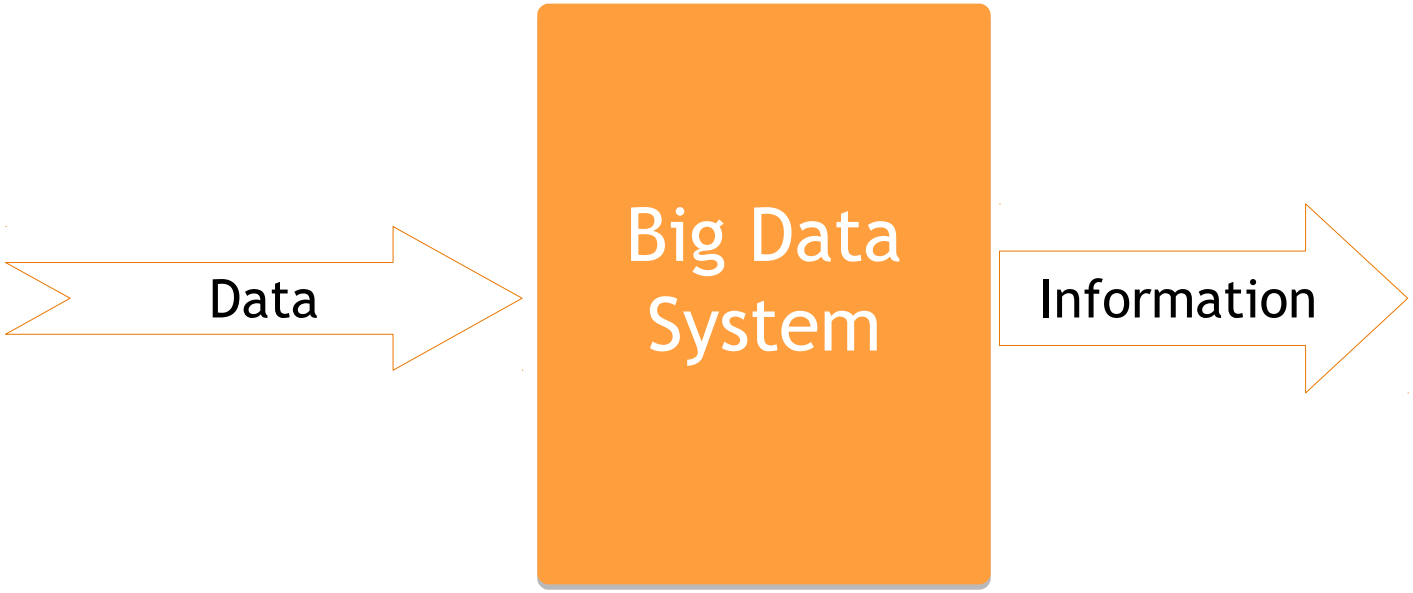
Система, работающая с Большими Данными

1. Система **работает** с данными!



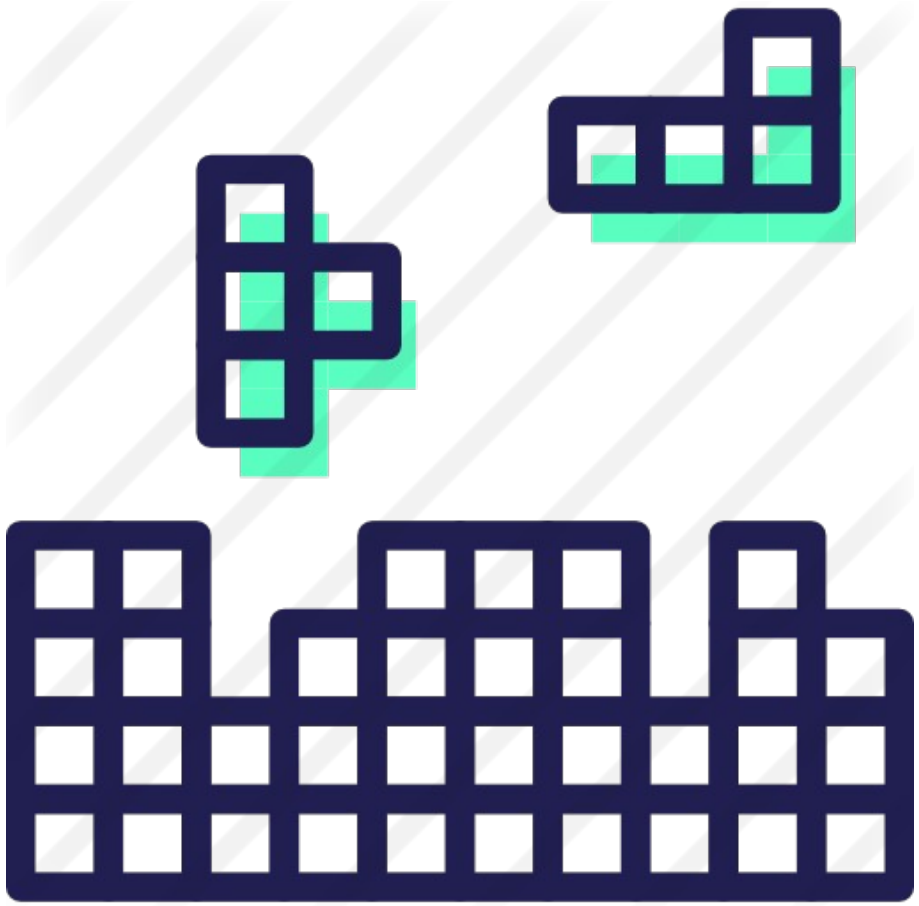
Система, работающая с Большими Данными

1. Система **работает** с данными!



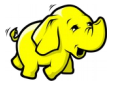
Система, работающая с Большими Данными

2. Состояние системы зависит от количества **обработанных** данных



Система, работающая с Большими Данными

3. Система использует “Big Data” технологии


Apache Hadoop


Apache Spark


Apache Storm


Apache Kafka


Apache Hive


Apache Pig


Graphite


Apache Flume


Apache HBase


Apache Solr


Apache Oozie


Apache Airflow


Elasticsearch


Aerospike

Система, работающая с Большими Данными

4. Данных “много”

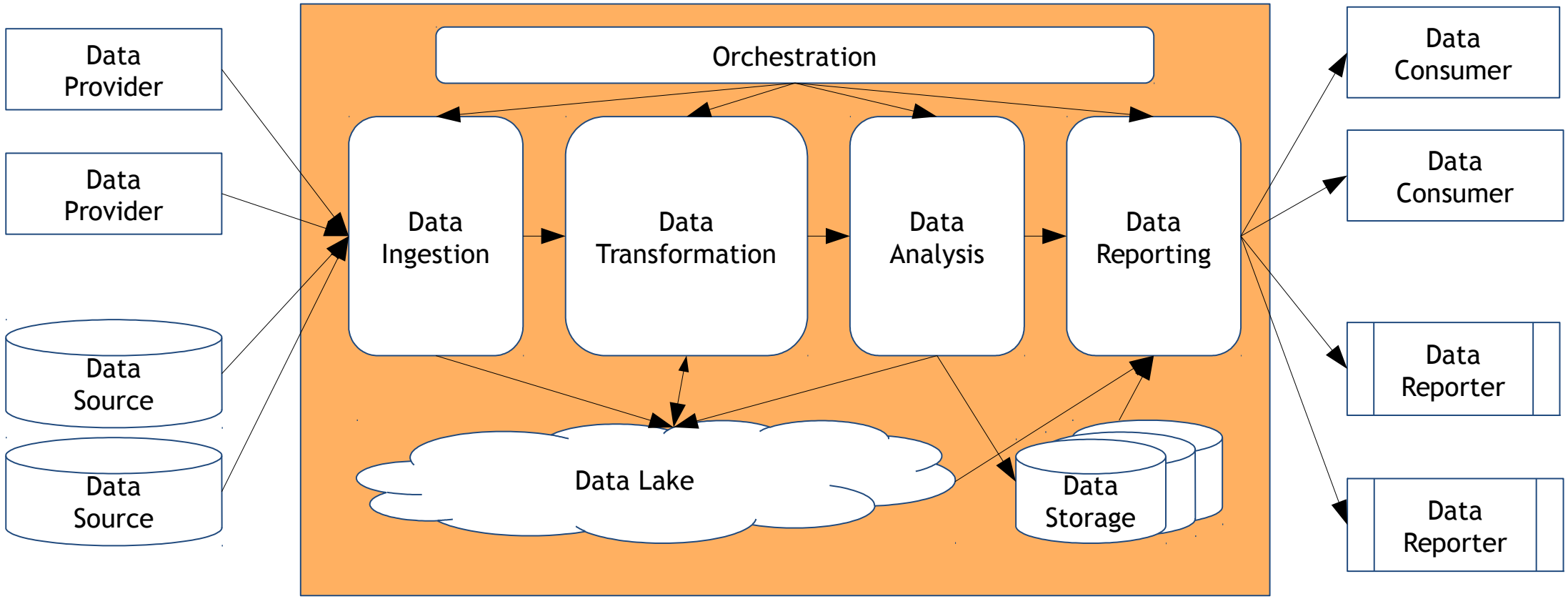
Система, работающая с Большими Данными

Это

1) Система **работает** с данными!
2) Состояние системы зависит от количества **обработанных** данных
3) Система использует “Big Data” технологии
4) Данных “много”

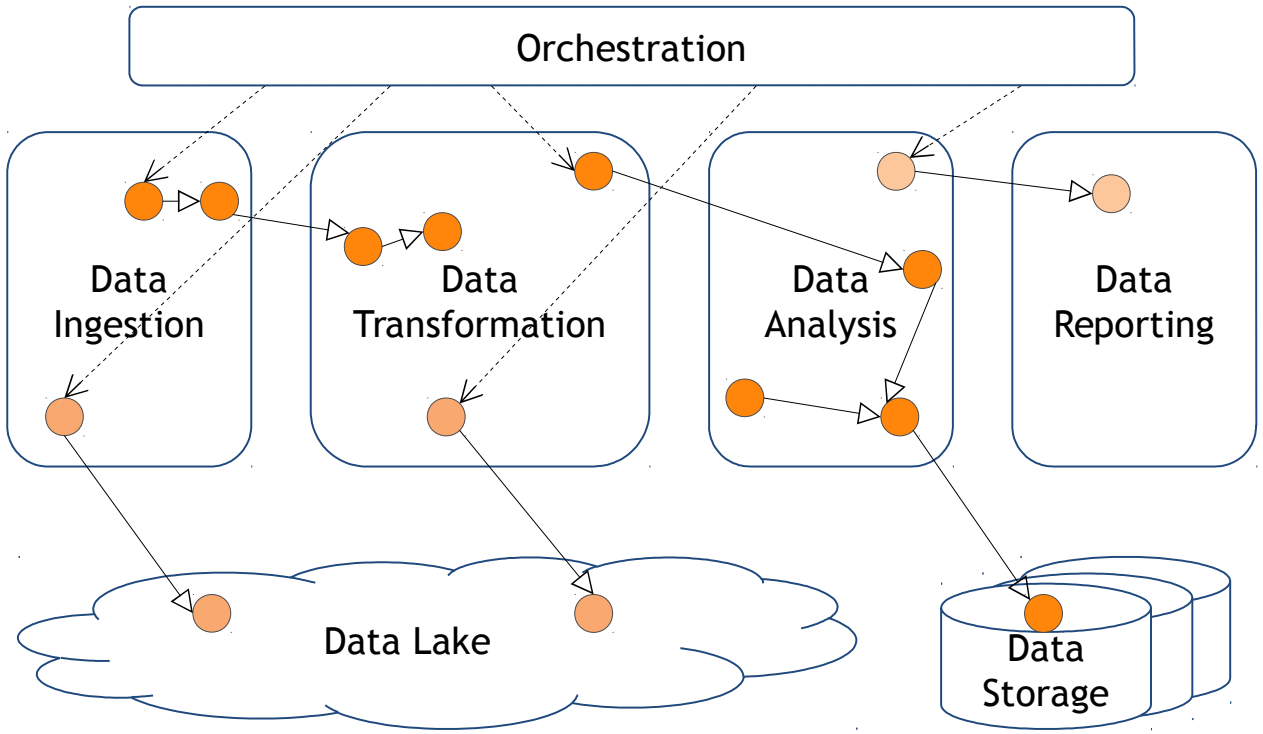
Система, работающая с Большими Данными

Обзор



Система, работающая с Большими Данными

Пайплайны и джобы



Система, работающая с Большими Данными

Терминология

Data Provider – the active component that pushes data into the system (e.g. data streams)

Data Source – the passive component from which the system fetches the data (e.g. DBMS)

Data Reporter – the active component which fetches the data from the system (e.g. clients' dashboards)

Data Consumer – the passive component to which the system pushes the data (e.g. partners' systems)

They are not part of the system

Data Storage – the part of the system that accumulates input, intermediate or resulting data

Data Lake – the special Data Storage to accumulate both raw and structured data

Data Ingestion – the process of the obtaining new data into the system and storing it in appropriate Data Storage

Data Transformation – the process of converting data from one format or structure into other one

Data Analysis – the process of evaluating data using analytical and logical approaches and algorithms

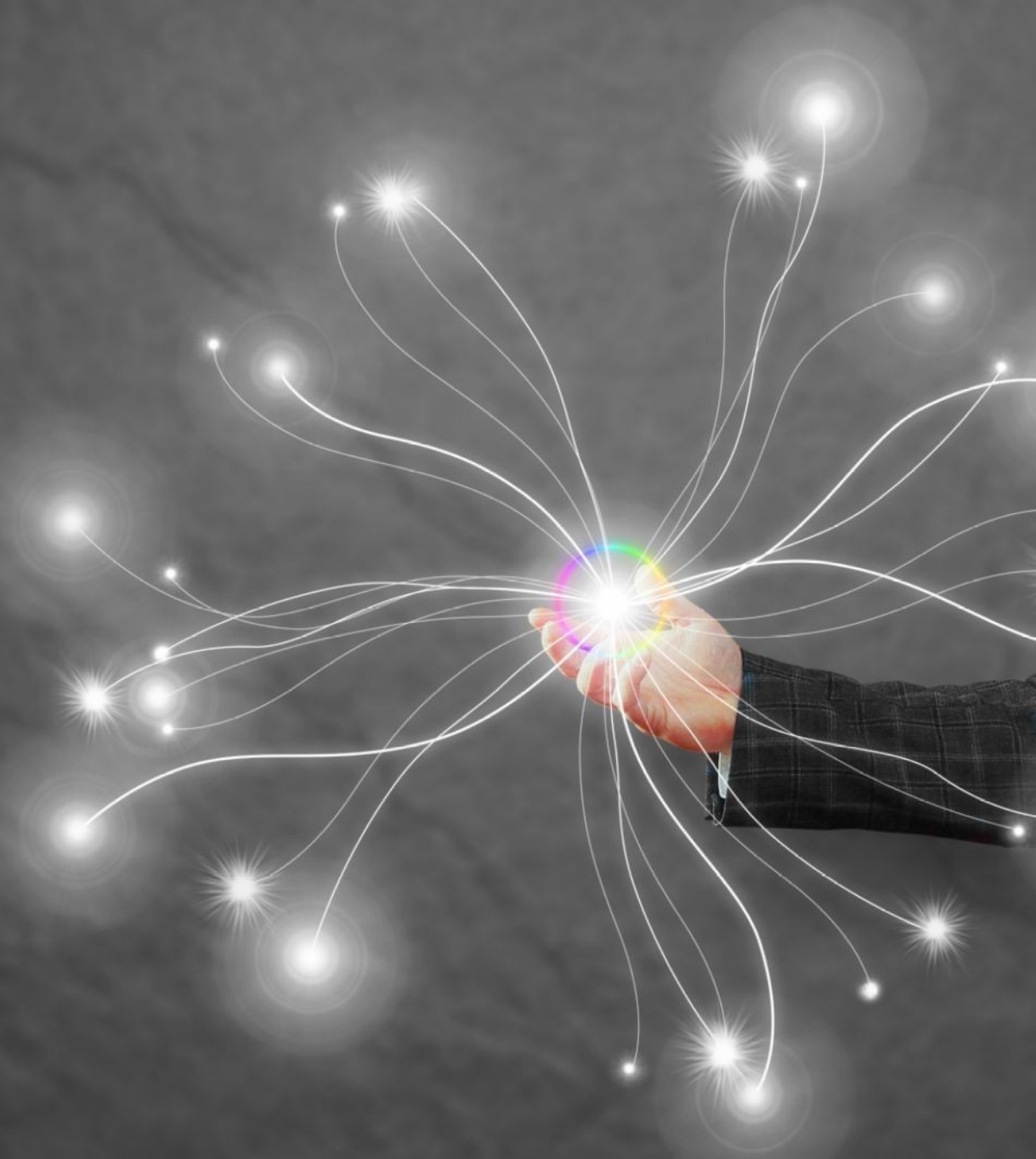
Data Reporting – the process of preparing and storing the output data to be accessed by external systems

Job – the single unit of work that requires input data and produces output data

Pipeline – the sequential set of jobs that processes the data where one job's output is the next job's input

Orchestration – the coordinated processing of multiple jobs, often with a conditional workflow

Контроль качества в Big Data системах



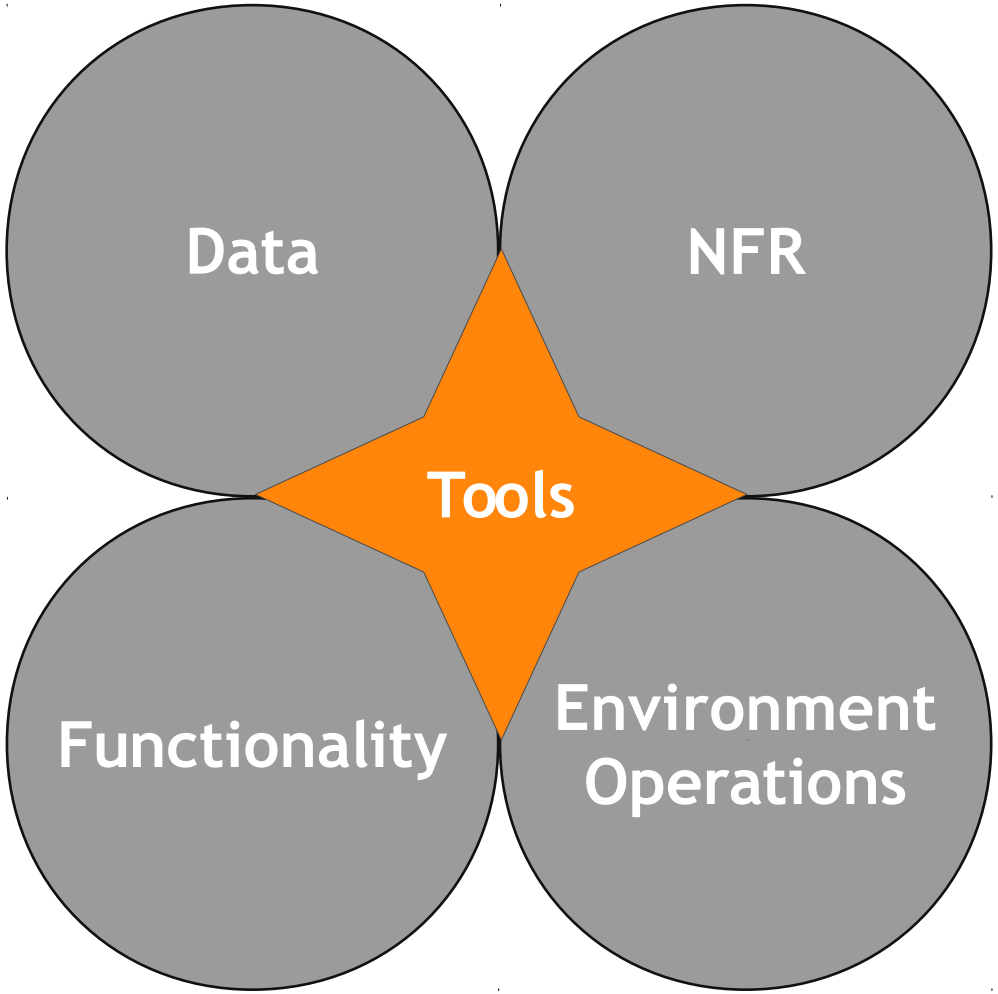
План

О чем будет данная презентация

- Что такое Big Data?
- **Контроль качества в Big Data системах**
  - Качество Данных
  - Функциональное Качество
  - Качество ”Стендов”
  - Нефункциональное Качество
  - Инструментарий
- Что нужно тестировщику, чтобы работать с Big Data

Контроль качества системы

Области



Качество Данных

Качество Данных

Обзор

Data Model Building

- Identification of
  - Data types
  - Data type's attributes and features

Test Data Management

- Sampling
- Generating
- Versioning

Data Certification

- Data Ingestion Control
- Output Data Checking

Качество Данных

Построение Модели Данных

Raw Data:

123.123.123.123 - - [13/Sep/2017:01:38:10 -0400] "GET /index.html HTTP/1.1" 200 - "-" "-"

Attributes:

IP: 123.123.123.123

Date: 13/Sep/2017:01:38:10 -0400

Method: Get

Path: /index.html

Status Code: 200

Features:

Country: China (from IP)

City: Beijing (IP)

Organization: China Unicom Beijing (IP)

Day of Week: Wednesday (Date)

Day Type: Workday (Date)

Time of Day (client): Night (Date)

Time of Day (server): Afternoon (Date)

Качество Данных

Основные ошибки в данных

- Данных полностью или частично нет в хранилище
- Данные ”поломанные” (нет ключевых атрибутов)
- Данные несогласованны
- Дубликаты
- Данные поступают в систему несвоевременно
- Данные не поступили в систему (например, обрыв сети)
- Изменился формат данных
- Данных приехало больше, чем ожидалось

Функциональное Качество

Функциональное Качество

Обзор

Data Flow Model Building

- Identification of dependencies between Job/Pipeline and Data Types and Data Storages

Testing Approaches

- Test Designing
- Test Levels

Automated Testing

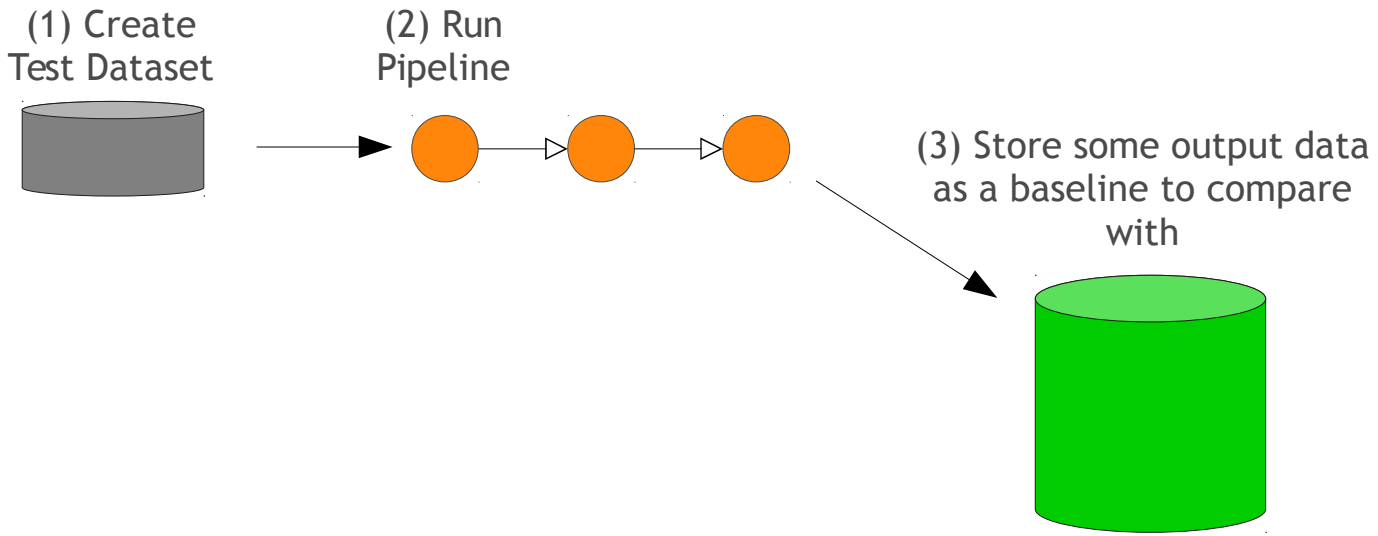
- Regression
- Orchestration
- API Testing

Change Impact Analysis

- Identifying change consequences
- Identifying conditions for reaching some goal

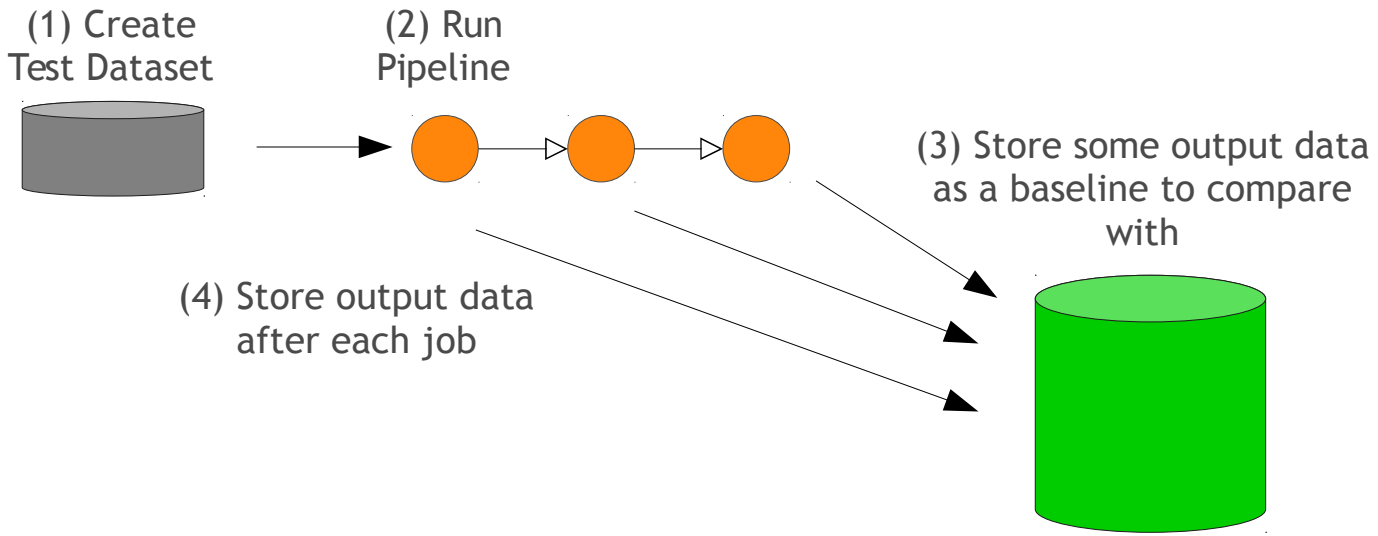
Функциональное Качество

Подход к интеграционному тестированию пайплайна



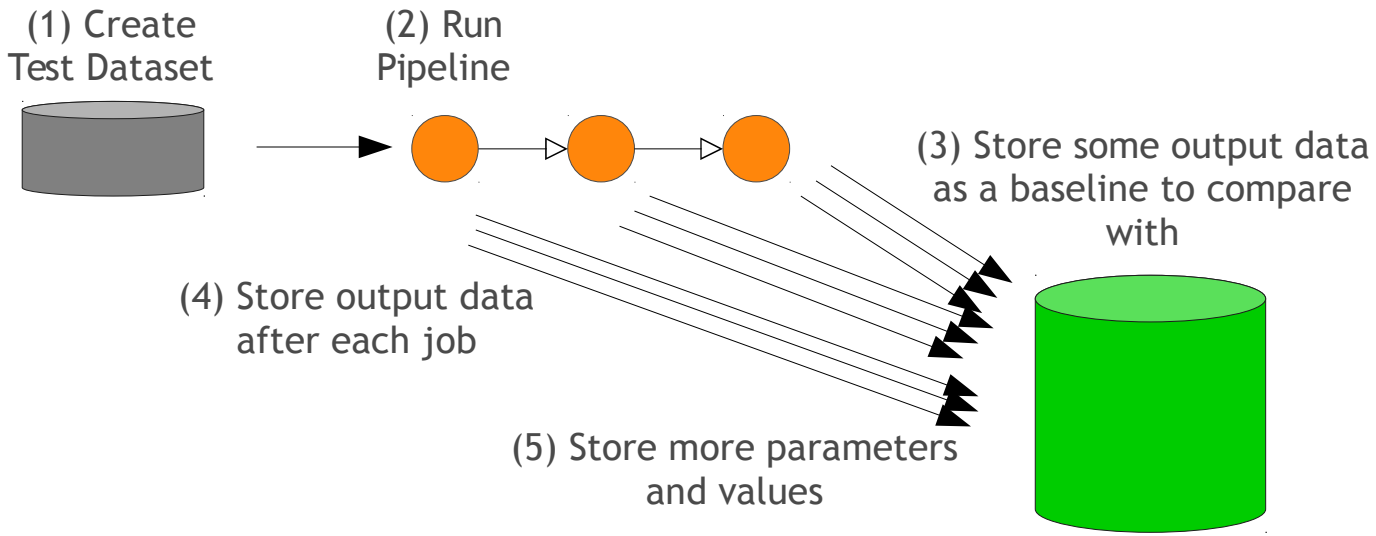
Функциональное Качество

Подход к интеграционному тестированию пайплайна



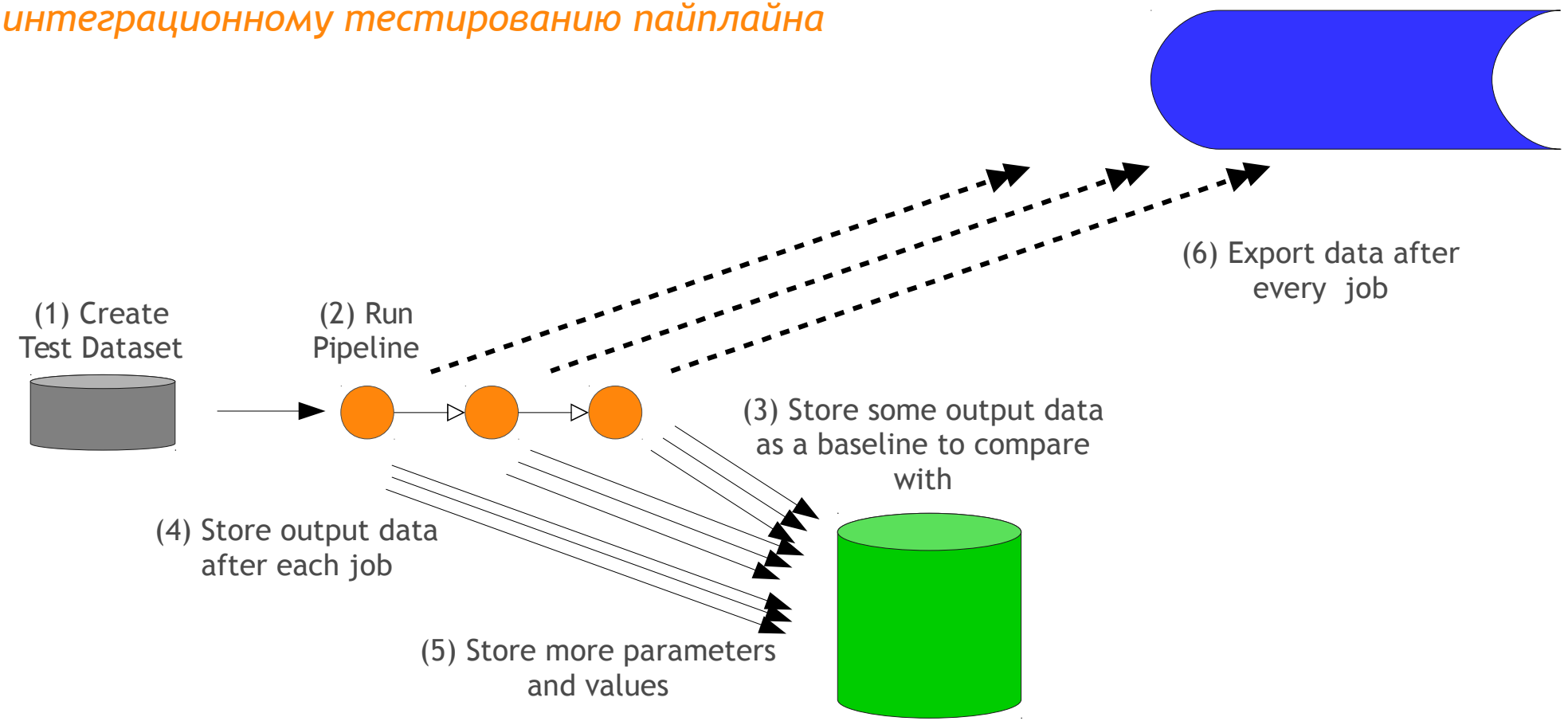
Функциональное Качество

Подход к интеграционному тестированию пайплайна



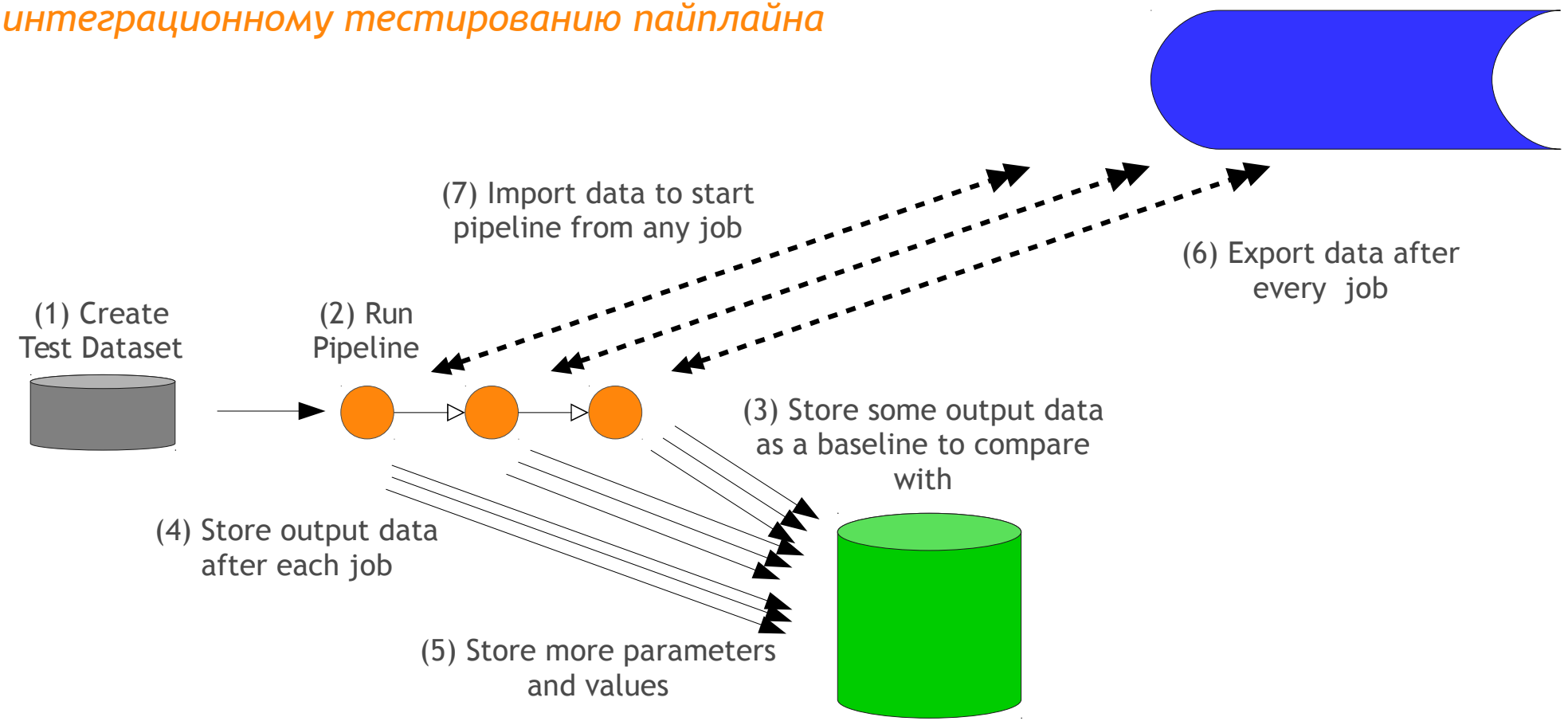
Функциональное Качество

Подход к интеграционному тестированию пайплайна



Функциональное Качество

Подход к интеграционному тестированию пайплайна



Качество Стендов

Качество Стендов

Обзор

Heath Checks

- Checking that all services are up and responding

Log Management

- Gathering
- Visualization
- Querying

Configuration Testing

- Testing that all properties are correct

Качество Стендов

Контроль конфигурации системы

1. Сделать единое хранилище пропертей.
2. Категоризировать проперти (static/runtime; environment/system; ...).
3. Определить дефолтные значения пропертей.
4. Продумать механизм переопределения пропертей.
5. Встроить в единую систему деплоймента.
6. Проверять конфигурацию **перед** деплойментом.
7. Лишние проперти, незаполненные параметры, разные значения одного и того же параметра, ссылки на отсутствующие *сервисы* или пути.

Нефункциональное Качество

Нефункциональное Качество

Обзор

Performance

- System performance
- Pipelines performance
- Profiling

Stability

- Robustness
- Scaling
- Endurance

Rollback / Back-out

- Testing that new changes could be rolled back

Нефункциональное Качество

Профилирование

(1) Add metrics framework

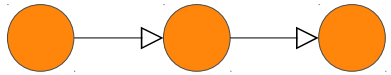
(2) Setup ELK

Нефункциональное Качество

Профилирование

(1) Add metrics framework

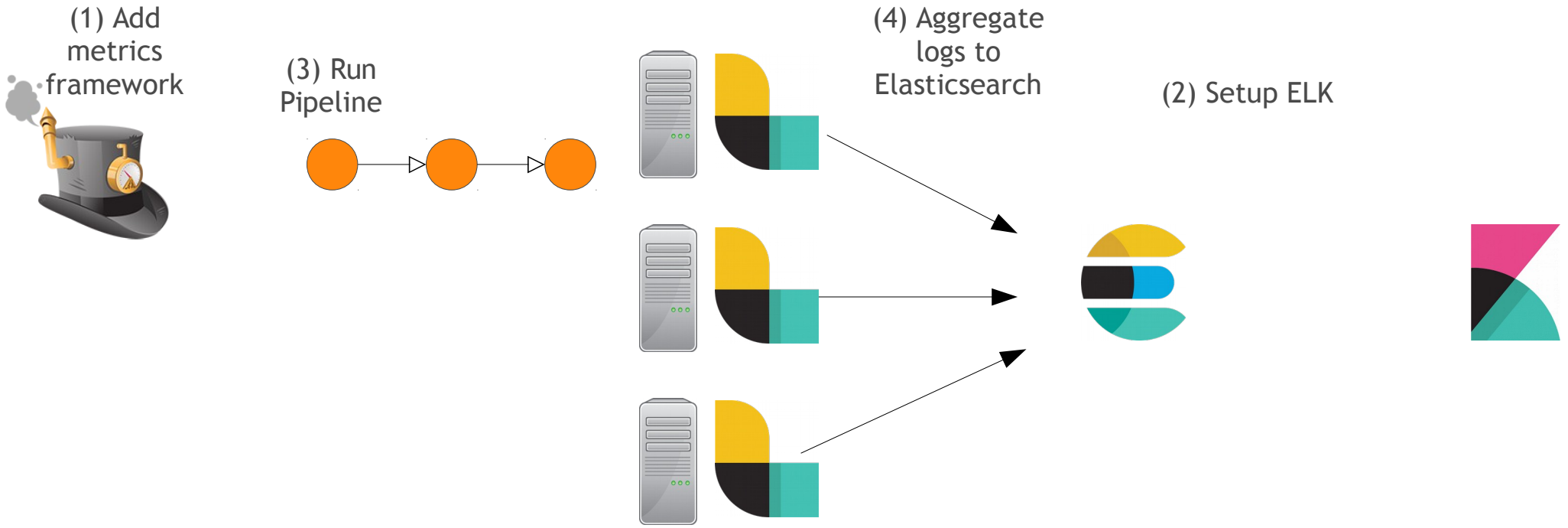
(3) Run Pipeline



(2) Setup ELK

Нефункциональное Качество

Профилирование

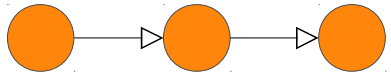


Нефункциональное Качество

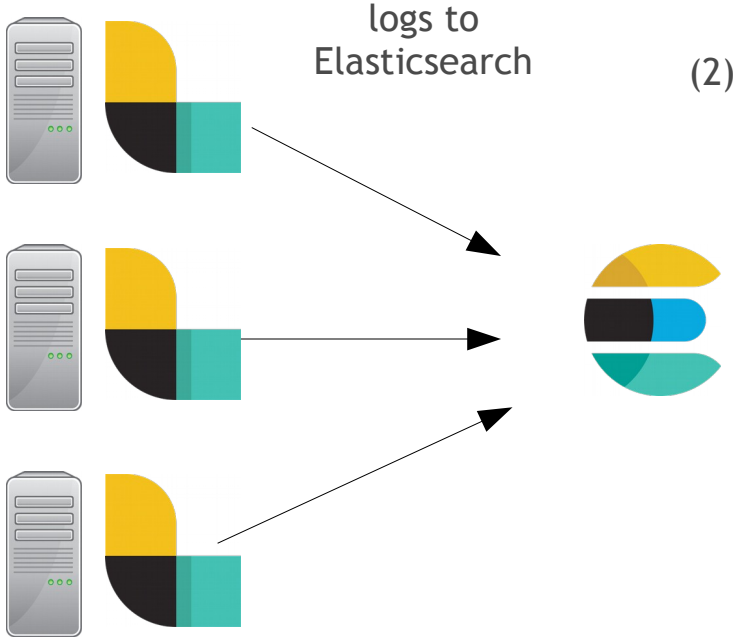
Профилирование

(1) Add metrics framework

(3) Run Pipeline

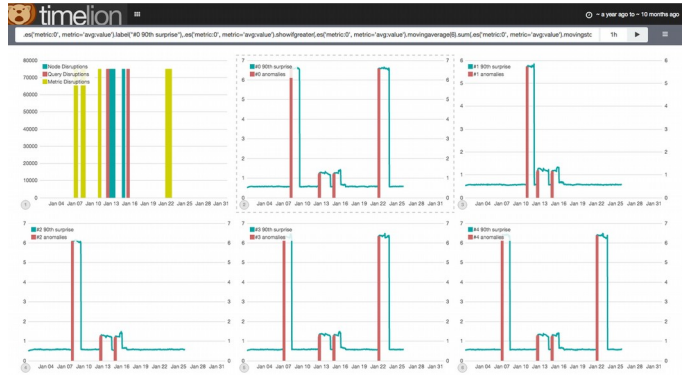


(4) Aggregate logs to Elasticsearch



(2) Setup ELK

(5) Visualize min/max/average for top 5 methods

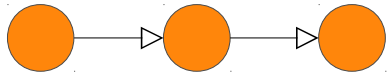


Нефункциональное Качество

Профилирование

(1) Add metrics framework

(3) Run Pipeline



(4) Aggregate logs to Elasticsearch

(2) Setup ELK

(5) Visualize min/max/average for top 5 methods

(6) Wait for fix and repeat until all fixed. Then monitor

Инструментарий

Инструментарий

Обзор

Data Management

- Generator
- Samplers
- Validators
- Analyzer

Statistics

- Gathering
- Visualization
- Querying

Test Frameworks

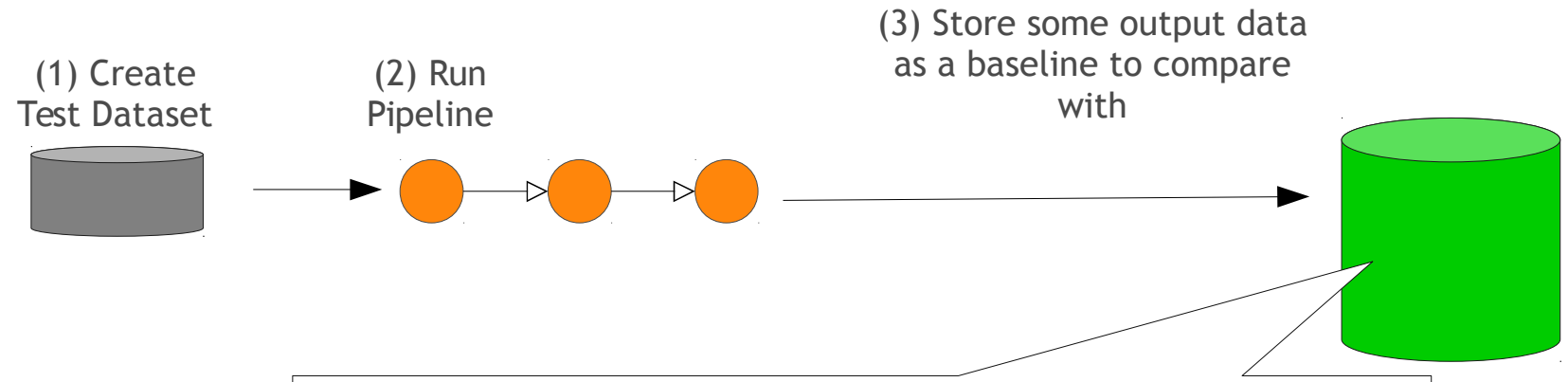
- Common
- Integrations with components

Environment Management

- Environment Constructor
- Log Management
- Monitoring

Инструментарий

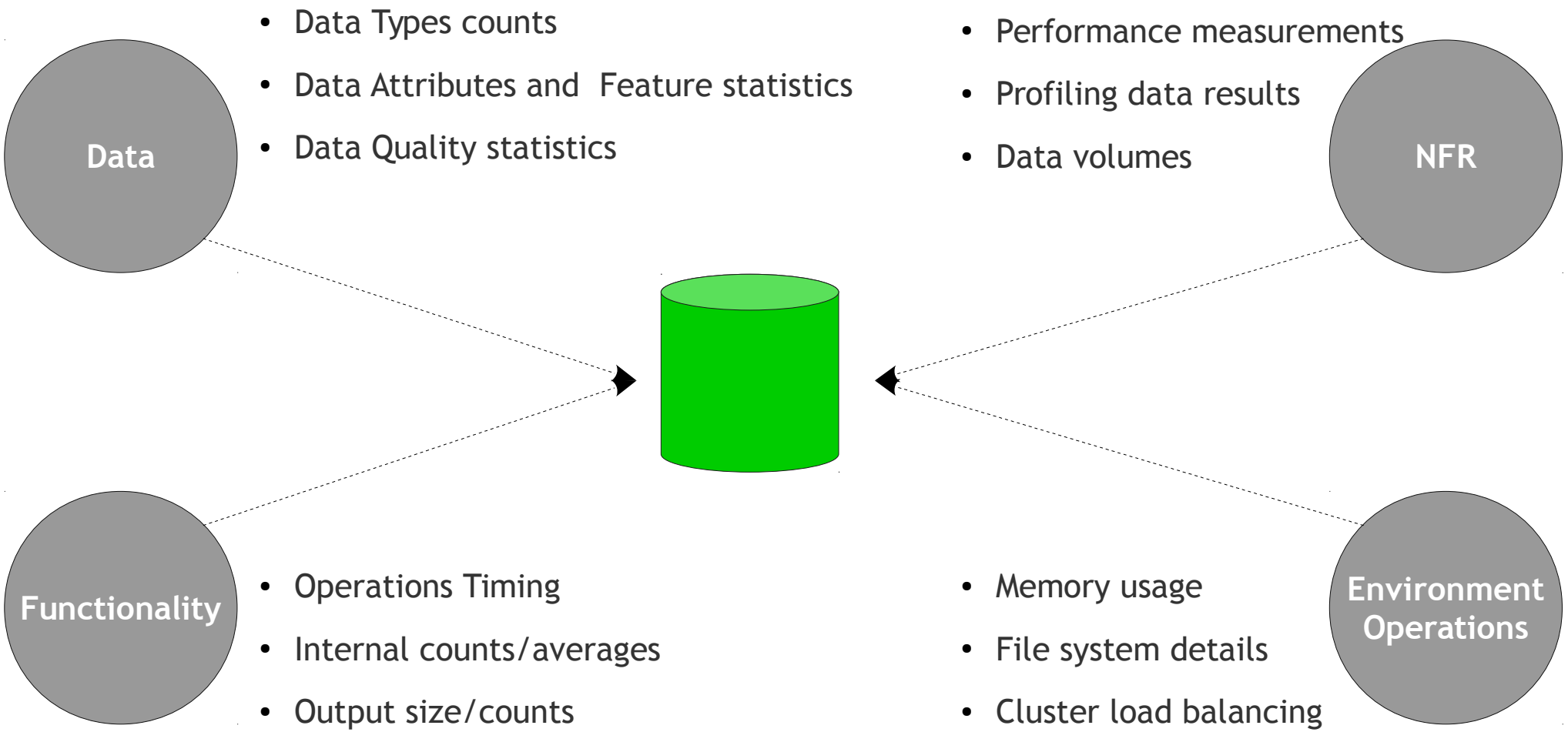
Статистика



Key	Run 1	Run 2	Run 3	Run 4	Run 5
Srch:index1:size	127	127	128	128	128
Srch:index2:size	10	10	12	12	13
HBase:tbl1:count	11122	11122	11122	11122	11122
HBase:tbl2:count	190	110	90	90	90
MySQL:tbl1:count	2834	2834	2834	2834	2333
Job1:time	9200	9480	9500	9430	9230
Log1:filesize	113	113	120	119	119

Инструментарий

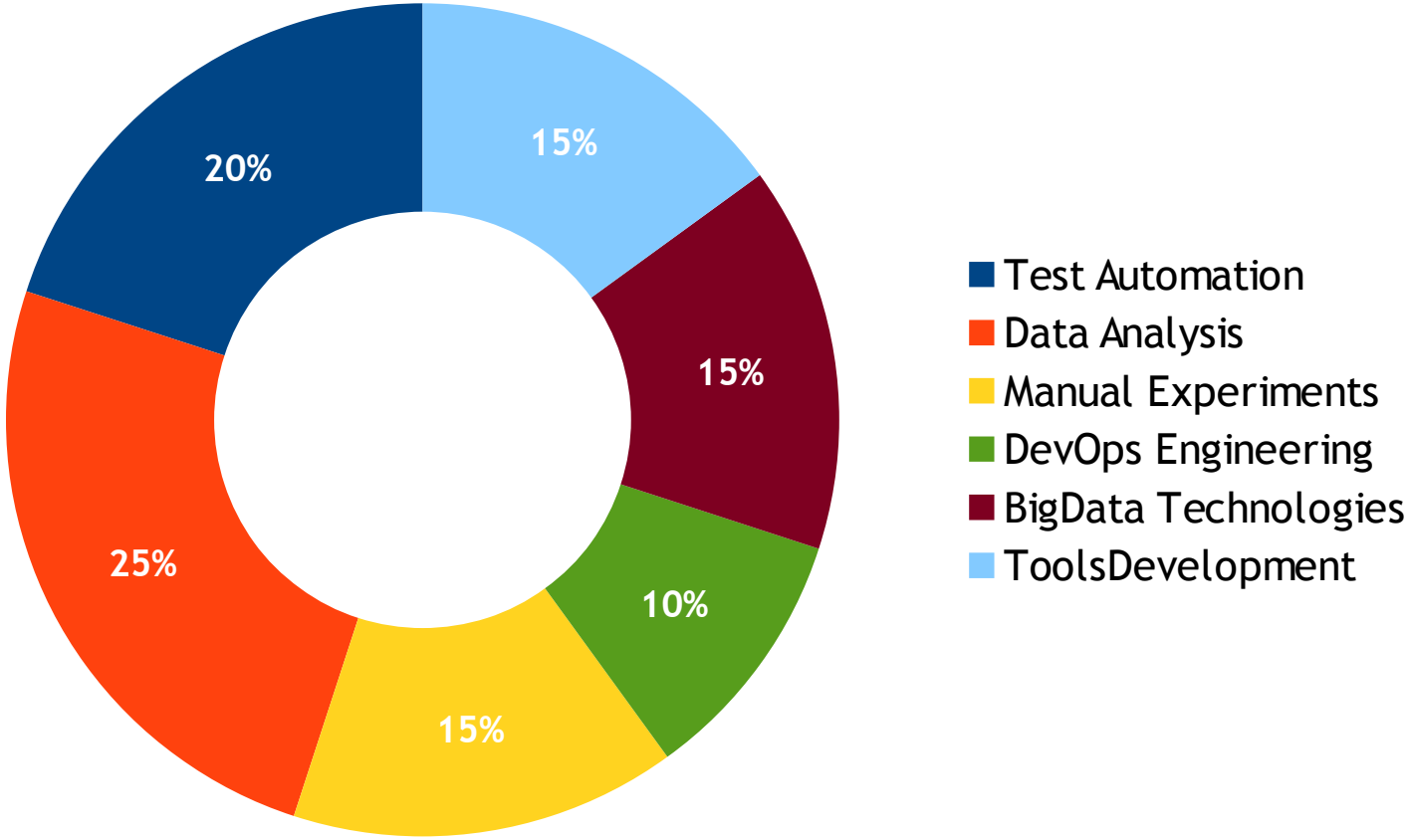
Статистика



Что нужно тестировщику, чтобы работать с Big Data

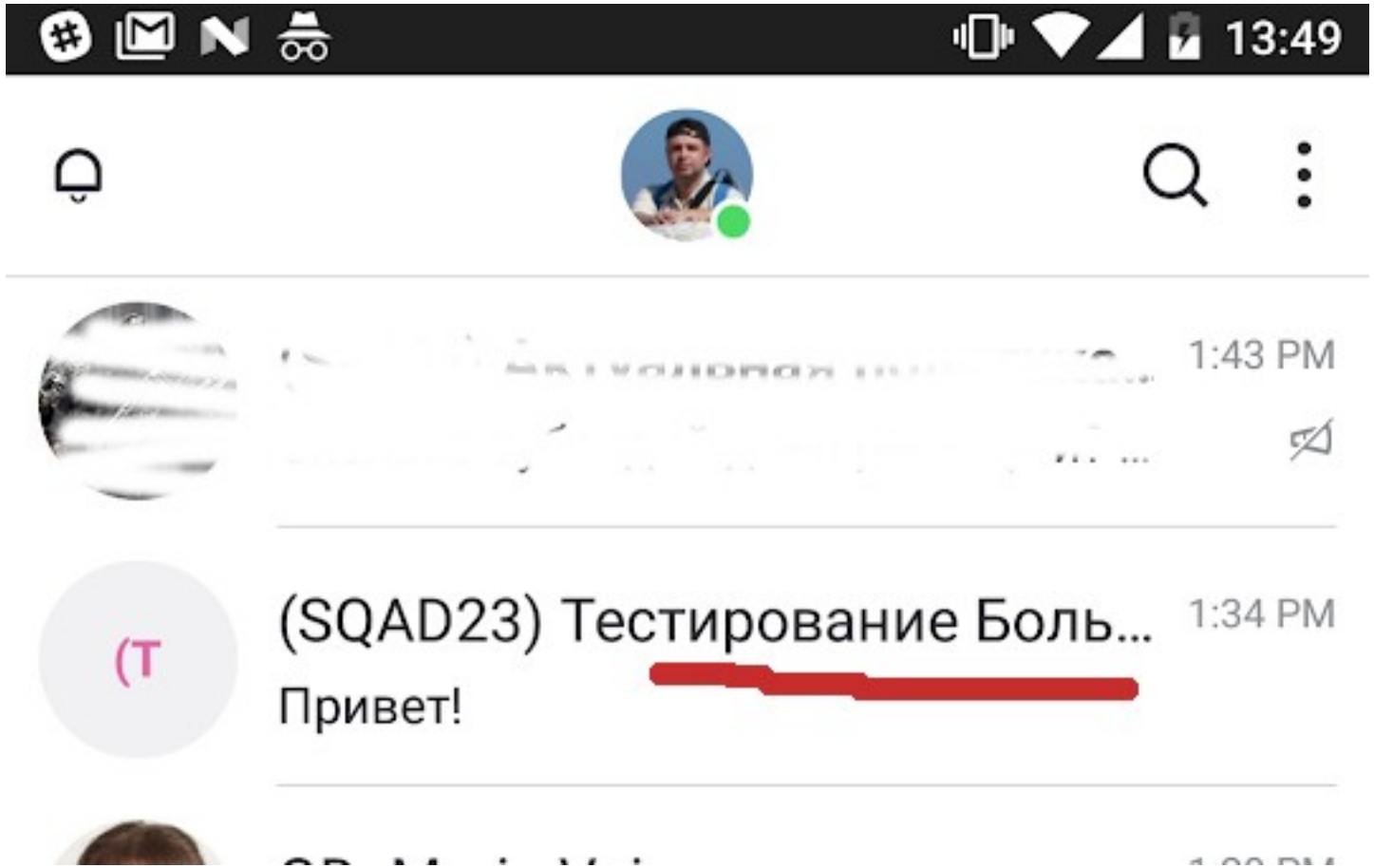
Распределение умений

Для команды тестирования



Тестирование Боль...

...ших Данных



SQA DAYS#23

Grid Dynamics

Спасибо!

www.griddynamics.com