informatica
Put potential to work.™

DIS

Переход от исследований к монетизации данных в промышленной эксплуатации

DIS Group

Петр Борисов,
Руководитель направления Big Data

«Я открыл монетизацию данных!!!»


«Оказывается, это не так то просто…»


Hadoop On-Premise
cloudera
Hortonworks
MAPR
Pivotal
IBM InfoSphere
splice MACHINE
bluedata
jethro
Hadoop in the Cloud
amazon web services
Microsoft Azure
Google Cloud Platform
IBM InfoSphere
CAZENA
TREASURE DATA
altiscale
Qubole
xplenty
Spark
databricks
GridGain
TACHYON NETWORKS
Cluster Services
amazon web services
kubernetes
HPCC SYSTEMS
docker
MESOSPHERE
CoreOS
pepperdata
StackIQ
Analyst Platforms
Palantir
AYASDI
Quid
enigma
Digital Reasoning
ORBITAL INSIGHT
Analytics Platforms
Microsoft
guavus
Datameer
interana
Data Science Platforms
context relevant
CONTINUUM ANALYTICS
DataRobot
Alpine
plotly
ADATAO
MODE
dataiku
Qutonian
DOMINO
sense
yhat
ALGORITHMIA
Visualization
tableau
Google Cloud Platform
Roambi
ZOOMDATA
Qlik
CHARTIO
Sales & Marketing
RADIUS
Gainsight
bloomreach
Zeta
livefyre
blue yonder
kahuna
Lattice
SAILTHRU
persado
infer
6sense
NG DATA
Wise.io
AVISO
ACTIONIQ
QUANTIFIND
NGAGIO
Customer Service
MEDALLIA
ATTENSITY
CLARABRIDGE
STELLAService
Preact
DigitalGenius
appuri
fuse machines
Human Capital
gild
Connectifier
textio
entelo
hiQ
Legal
RAVEL
JUDICATA
Everlaw
Brevia
PREMONITION
NoSQL Databases
amazon DynamoDB
Google Cloud Platform
ORACLE
Microsoft Azure
MarkLogic
mongoDB
DATASTAX
AEROSPIKE
Couchbase
SequoiaDB
redislabs
influxdata
NewSQL Databases
SAP HANA
Clustrix
Pivotal
paradigm4
memsql
NUODB
MariaDB
VOLTDB
citusdata
deepdb
Trafodion
Cockroach LABS
BI Platforms
Power BI
amazon web services
DOMO
salesforce
Wave Analytics
GoodData
birst
platfora
looker
atscale
ARCADIA DATA
SISENSE
Statistical Computing
sas
SPSS
MATLAB
Log Analytics
splunk>
sumologic
kibana
CLOUD PHYSICS
loggly
Social Analytics
NETBASE
DATASIFT
tracx
bitly
synthesio
bottlenose
simplereach
Ad Optimization
MediaMath
Integral Ad Science
rocketfuel
OpenX
theTradeDesk
Adgorithms
LiveIntent
dstillery
DataXu
Appier
TAPAD
Security
CYLANCE
CounterTack
cybereason
ThreatMetrix
AREA 1 SECURITY
SentinelOne
Guardian Analytics
Recorded Future
FORTSCALE
sift science
Keybase
feedzai
SIGNIFYD
Vertical AI Applications
facebook
Clara
KASISTO
lumiata
Graph Databases
neo4j
GIRAPH
OrientDB
InfiniteGraph
MPP Databases
TERADATA
VERTICA
An HP Company
NETEZZA
An IBM Company
kognitio
dremio
Cloud EDW
amazon web services
Google Cloud Platform
Microsoft Azure
Pivotal
snowflake
WATERLINE DATA
Infoworks
Data Transformation
alteryx
TRIFACTA
tamr
Paxata
StreamSets
Alation
Data Integration
informatica
Put potential to work:
MuleSoft
snapLogic
BedrockData
Real-Time
amazon web services
METAMARKETS
confluent
DATATORRENT
dataArtisans
Machine Learning
Azure Machine Learning
amazon web services
H2O.ai
Dato
SKYTREE
rapidminer
DATARPM
deepsense.io
VISENZE
PredictionIO
glowfish
Speech & NLP
NarrativeScience
api.ai
NUANCE
Gridspace
semantic machines
cortical.io
MindMeld
IDIBON
vsevo
Horizontal AI
IBM Watson
Cortana
sentient technologies
VIV
vicarious
nervana SYSTEMS
Numenta
nara
SCALED INFERENCE
HyperScience
MetaMind
clarifai
DEXTRO
Geometric Intelligence
Publisher Tools
outbrain
mixpanel
Chartbeat
yieldbot
Yieldmo
Govt/ Regulation
Socrata
OPENGOV
FiscalNote
enigma
PREDPOL
mark43
OpenDataSoft
Finance
Affirm
LendingClub
OnDeck
Kreditech
zestfinance
LendUp
Kabbage
tidemark
Payoff
INSIKT
uora
Dataminr
Lenddo
KENSHO
AIDYIA
iSENTIUM
Quantopian
sentient technologies
Management / Monitoring
New Relic.
APPDYNAMICS
amazon web services
actifio
Numerify
splunk
DATADOG
rocana
Anodot
Security
TANIUM
illumio
CODE42
DataGravity
CipherCloud
VECTRA
sqrrl
BlueTalon
Storage
amazon web services
Google Cloud Platform
Microsoft Azure
panasas
nimblestorage
Qumulo
App Dev
apigee
CASK
Keen IO
Typesafe
CONCURRENT
Crowd-sourcing
amazon mechanical turk
CrowdFlower
WorkFusion
Search
Autonomy
ORACLE
ENDECA
EXALEAD
Lucidworks
elastic
ThoughtSpot
MAANA
swiftype
Algolia
SINEQUA
Data Services
Mu Sigma
OPERA SOLUTIONS
DATA SCIENCE
SILICON VALLEY DATA SCIENCE
kaggle
datascope
DataKind
For Business Analysts
OrigamiLogic
ClearStory
CIRRO
import.io
SMB / Commerce
Google Analytics
AMPLITUDE
RJMetrics
BLUECORE
sumall
granify
Airtable
retention SCIENCE
custora
Education/ Learning
KNEWTON
Clever
declara
PANORAMA
knowre
Life Sciences
23andMe
PATHWAY GENOMICS
Counsyl
deep genomics
Recombine
KYRUUS
FLATIRON
zymergen
HealthTap
METABIOTA
ZEPHYR HEALTH
ovia
Ginger.io
transcriptic
Glow
enlitic
AiCure
Atomwise
Industries
OPOWER
eHarmony
RetailNext
duetto
STITCH FIX
WorkFusion
BLUE RIVER
TACHYUS
SwiftKey
SeeQ
FarmLogs
HowGood
celect
SIGHT MACHINE
statmuse
BOXEVER
amazon web services
Google
Microsoft
IBM
SAP
sas
Autonomy
vmware
talend
TIBCO
TERADATA
ORACLE
NetApp
Framework
hadoop HDFS
hadoop MapReduce
YARN
Spark
MESOS
TEZ
Flink
CDAP
Apache Kylin
Query / Data Flow
SLAMDATA
HIVE
APACHE DRILL
Google Cloud Dataflow
Data Access
cassandra
APACHE HBASE
accumulo
mongoDB
kafka
SciDB
nifi
CouchDB
riak
OPENTSDB
Coordination
talend
ZOOKEEPER
Apache Zookeeper
Apache Ambari
Real-Time
STORM
Spark
APEX
Flink
TACHYON
druid
Stat Tools
R
Scala
NumPy
SciPy
Machine Learning
mllib
mahout
Apache SINGA
MADlib
Aerosolve
Caffe
learn
torch
CNTK
TensorFlow
FeatureFu
WEKA
jupyter
DL4J
Search
elasticsearch
Solr
Security
Apache Ranger
Visualization

```
public class LookupReducer extends Reducer<TextPair,Text,T

private String result = "";
private String msisdn;
private String attribute, product;
private long trans_dt_long, start_dt_long, end_dt_long;
private String trans_dt, start_dt, end_dt;

@Override
public void reduce(TextPair key, Iterable<Text> values, Co
        throws IOException, InterruptedException {

     context.progress();
    //value without key to remember

    Iterator<Text> iter =  values.iterator();

 for (Text val : values) {

Text recordNoKey = val;     //new Text(iter.next());

String valSplitted[] = recordNoKey.toString().split(";");

//if the input is coming from CDR set corresponding values

    if(key.getSecond().toString().equals(CDR.CDR_TAG))
    {
        trans_dt = recordNoKey.toString();
        trans_dt_long = dateToLong(recordNoKey.toString())
    }
  //if the input is coming from Attributes set correspondi
    else if(key.getSecond().toString().equals(Attribute.AT
    {
        attribute = valSplitted[0];
        product = valSplitted[1];
        start_dt = valSplitted[2];
        start_dt_long = dateToLong(valSplitted[2]);
        end_dt = valSplitted[3];
        end_dt_long = dateToLong(valSplitted[3]);;
    }

        Text record = val;  //iter.next();
        //System.out.println("RECORD: " + record);
        Text outValue = new Text(recordNoKey.toString() +
```

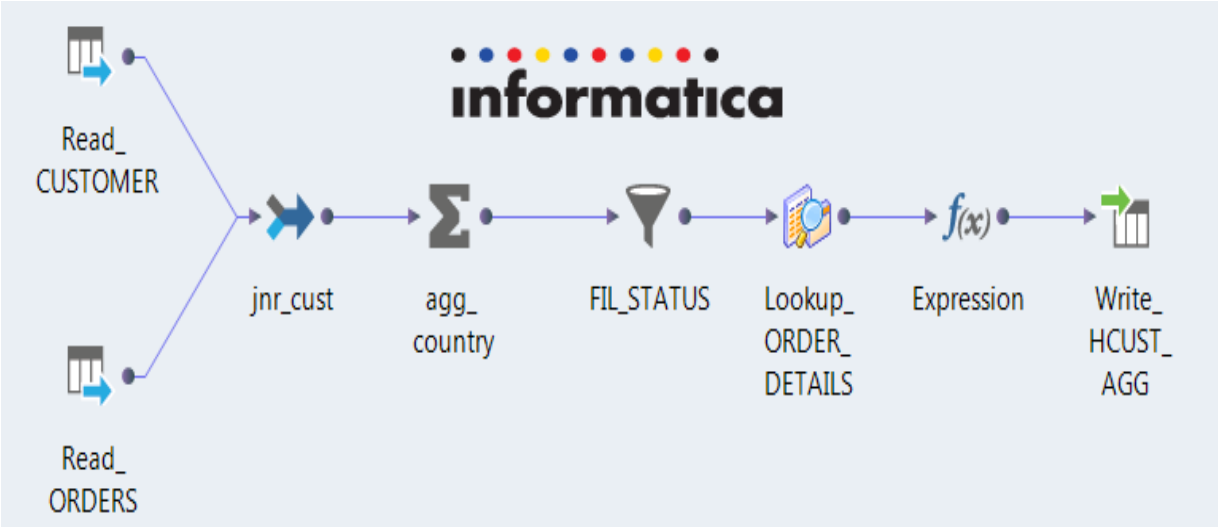
```
if(start_dt_long < trans_dt_long && trans_dt_long < end
    {
    //concat output columns
        result = attribute + ";" + product + ";" + tran

    context.write(key.getFirst(), new Text(result));
    System.out.println("KEY: " + key);
        }
    }
}

private static long dateToLong(String date){
    DateFormat formatter = new SimpleDateFormat("yyyy-M
    Date parsedDate = null;
    try {
        parsedDate = formatter.parse(date);
    } catch (ParseException e) {
        // TODO Auto-generated catch block
        e.printStackTrace();
    }
    long dateInLong = parsedDate.getTime();

    return dateInLong;

public static class TextPair implements WritableCompara

    private Text first;
    private Text second;

    public TextPair(){
        set(new Text(), new Text());
    }

    public TextPair(String first, String second){
        set(new Text(first), new Text(second));
    }

    public TextPair(Text first, Text second){
        set(first, second);
    }

    public void set(Text first, Text second){
        this.first = first;
        this.second = second;
```

```
public Text getFirst() {
    return first;
}

public void setFirst(Text first) {
    this.first = first;
}

public Text getSecond() {
    return second;
}

public void setSecond(Text second) {
    this.second = second;
}

@Override
public void readFields(DataInput in) throws IOException {
    // TODO Auto-generated method stub
    first.readFields(in);
    second.readFields(in);
}

@Override
public void write(DataOutput out) throws IOException {
    // TODO Auto-generated method stub
    first.write(out);
    second.write(out);
}

@Override
public int hashCode(){
    return first.hashCode() * 163 + second.hashCode();
}

@Override
public boolean equals(Object o){
    if(o instanceof TextPair)
    {
        TextPair tp = (TextPair) o;
        return first.equals(tp.first) && second.equals(tp.second);
    }
    return false;
}

@Override
```

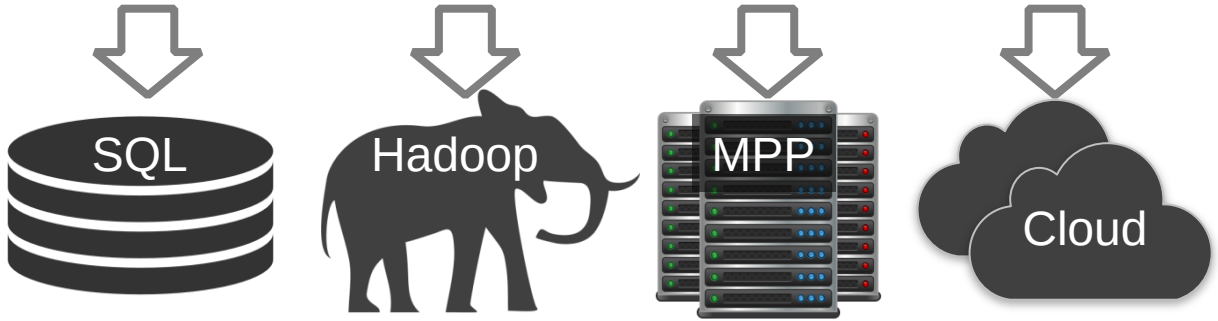
Универсальная платформа управления данными

DIS

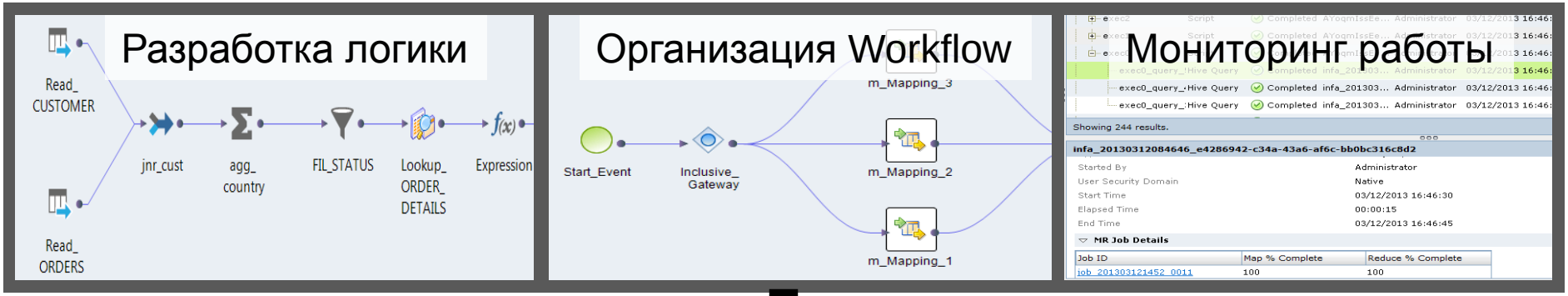
- Data Connectivity
- Data Integration
- Data Quality
- Data Mastering



- Self Service
- Data Governance
- Data Security
- Data Preparation



Использование Polyglot Engine



Smart Executor

Native

Native Informatica Data Transformation Engine

Кластер Hadoop

Hive on Map Reduce | Hive On Tez | Spark | Blaze

Map Reduce | Tez | Spark Core | Cluster Aware

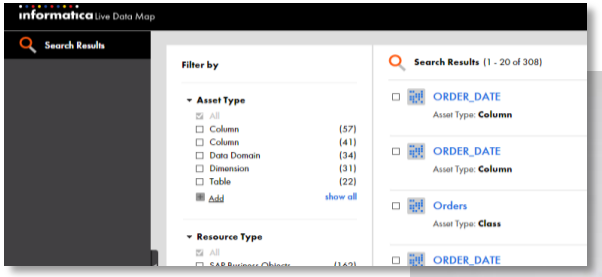
YARN

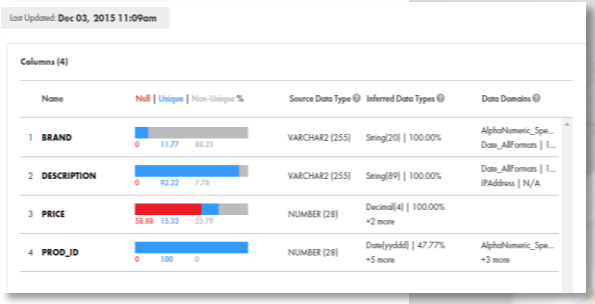
HDFS

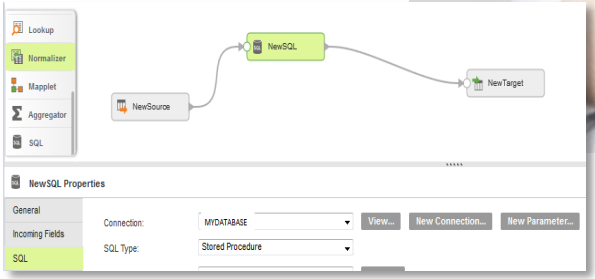
SQL

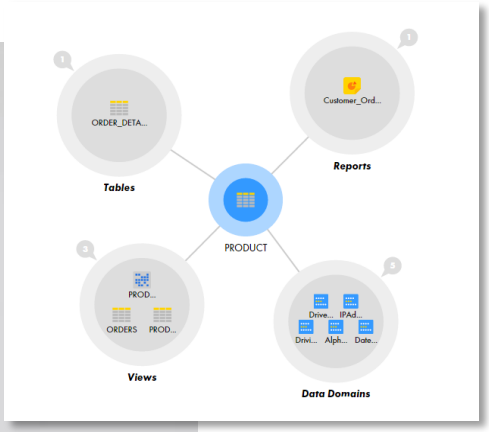
Database Pushdown

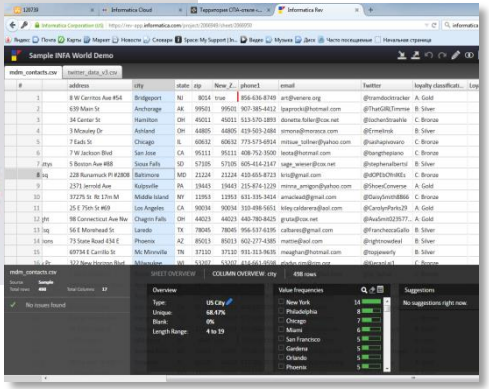
Self Service для пользователей





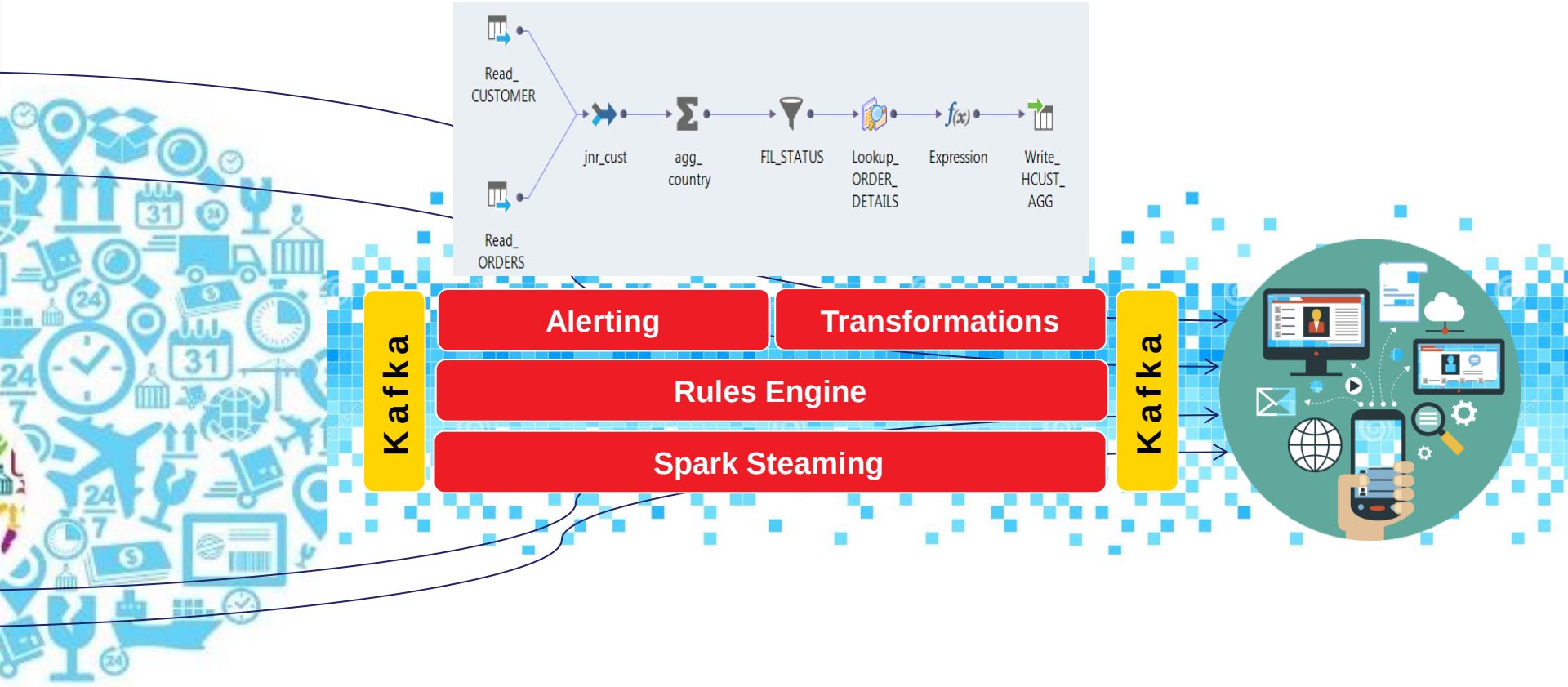






Потоковая онлайн-аналитика

DIS



Что дает Informatica в проектах Big Data

#1 Time-To-Market

- Скорость найма
- Скорость разработки
- Скорость изменений
- Вовлечение бизнеса

Что дает Informatica в проектах Big Data

#2 Надежность

- Снижение рисков
- Полный контроль
- Поддержка
- Возможность развития


informatica

Спасибо за внимание!
Ваши вопросы

DIS Group
Петр Борисов
info@dis-group.ru
+7 495 645 02 01