

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...»



This image is a comprehensive grid of logos for various technology companies, organized into categories. The categories are arranged in a grid, with some larger categories spanning multiple rows or columns.

Categories and Companies:

- On-Premise / Cloud:** 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 NEXUS), Cluster Services (amazon web services, kubernetes, HPC SYSTEMS, docker, MESOSPHERE, CoreOS, pepperdata, StackIQ), Analyst Platforms (Palantir, AYASDI, Quid enigma, Digital Reasoning, ORBITAL INSIGHT), Analytics Platforms (Microsoft, guAVUS, Datameer, inter|ana), Data Science Platforms (context relevant, CONTINUUM ANALYTICS, DataRobot, Alpine, MODE, dataku, tonian, DOMINO, sense, what, ALGORITHMIA), Visualization (tableau, Google Cloud Platform, Roambi, QOMDATA, Qlik, CHARTIO), Sales & Marketing (RADIUS, Gainsight, bloomreach, Zeta, livefyre, blueyonder, kahuna, Lattice, SAILTHRU, persado, infer, bsense, AVISO, ACTIONIQ, QUANTIFIND, ENGA GIO), Customer Service (MEDALLIA, ATTENSY, CLARABRIDGE, STELLA Service, NGDATA, Preact, DigitalGenius, appuri, fuse:machines), Human Capital (gild, Connectifier, textic, entelo, hiQ), Legal (RAVEL, JUDICATA, Everlaw, Brevia, PREM-ONITION).
- Databases:** NoSQL Databases (amazon DynamoDB, Google Cloud Platform, Microsoft Azure, ORACLE, mongoDB, MarkLogic, DATASTAX, Couchbase, SequoiaDB, redislabs, Influxdata), NewSQL Databases (SAP HANA, Clustrix, Pivotal, paradigm4, memsql, data-driven discovery, NUODB, MariaDB, VOLTDB, citusdata, deepdb, Trafodion, Cockroach LABS), BI Platforms (Power BI, amazon web services, Domo, Wave Analytics, GoodData, birst, FVVO Insights, platforma, looker, atscale, ARCADIA, SIBSENSE), Statistical Computing (sas, SPSS, MATLAB), Log Analytics (splunk, sumologic, kibana, CLOUD PHYSICS, loggly), Social Analytics (NETBASE, DATASIFT, tracx, bitly, synthetio, bottlenose, simple reach).
- Ad Optimization:** MediaMath, Integral Ad Science, OpenX, rocketfuel, Adgorithms, theTradeDesk, Liventent, dstillery, DataXu, Appier, TAPAD.
- Security:** CYLANCE, CounterTack, cybereason, ThreatMetrix, AREA 1 SECURITY, SentinelOne, Recorded Future, Guardian Analytics, FORTSCALE, sift science, Kaybase, feedzai, SICNIFYD.
- Vertical AI Applications:** facebook, Clara, KASIST, lumiata.
- Graph Databases:** neo4j, TERADATA, VERTICA, Netezza, OrientDB, InfiniteGraph, dremio, kognitio.
- MPP Databases:** amazon web services, Microsoft Azure, Pivotal, snowflake, WATERLINE DATA, Infoworks.
- Cloud EDW:** amazon web services, Microsoft Azure, Pivotal, snowflake, WATERLINE DATA, Infoworks.
- Data Transformation:** alteryx, TRIFACTA, tamr, Paxata, StreamSets, Alation.
- Data Integration:** informatica, MuleSoft, snapLogic, BedrockData.
- Real-Time:** amazon web services, METAMARKETS, confluent, DATATORRENT, dataArtisans.
- Machine Learning:** Azure Machine Learning, H2O, SKYTREE, rapidminer, DATARPM, deepsense, VISENZE, PredictionIO, glowfish, IDIBON, yscope.
- Speech & NLP:** NarrativeScience, api.ai, NUANCE, Gridspaces, semanticmachines, cortical.io, MindMeld, IDIBON, yscope.
- Horizontal AI:** IBM Watson, Cortana, sentient, VIV, nervana, nora, Numenta, MetaMind, clarifai, DEXTR0, Geometric Intelligence.
- Publisher Tools:** outbrain, mixpanel, Chartbeat, yieldbot, Yieldmo.
- Govt/ Regulation:** Socrata, OPENGOV, EN FiscalNote, enigma, PREPOL, mark43, OpenDataSoft.
- Finance:** affirm, LendingClub, OnDeck, Kreditech, zest finance, LendUp, Kabbage, tidemark, Playfit, INSIKT, ZUORA, Dataminr, Lenddo, KENSHC, AIDYIA, ISENTIUM, Quantopian, sentient technology.
- Management / Monitoring:** New Relic, APPDYNAMICS, amazon web services, octifio, Numerify, splunk, DATADOG, Trocana, Anodot.
- Security:** TANIUM, illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrrl, BlueTalon.
- Storage:** amazon web services, Microsoft Azure, panasas, nimblestorage, Qumulo.
- App Dev:** apigee, CASK, Keren IO, Typesafe, CONCURRENT.
- Crowd-sourcing:** amazon mechanical turk, CrowdFlower, WorkFusion.
- Search:** hp, Autonomy, ORACLE ENDECA, EXALSER, Lucidworks, elastic, ThoughtSpot, MAANA, swifttype, Algolia, SINEQUA.
- Data Services:** OPERA, MU SIGMA, DATA SCIENCE, OPERA, DATA SCIENCE, kaggle, datascopes, DataKind.
- For Business Analysts:** OrigamiLogic, ClearStory, CIRRO, import IO.
- SMB / Commerce:** Google Analytics, AMPLITUDE, RJMetrics, BLUECORE, sumall, granify, Airtable, retention, custora.
- Education/ Learning:** KNEWTON, Clever, deClara, PANORAMA, knowre.
- Life Sciences:** 23andMe, Pathway Genomics, Recombine, deep genomics, KYRUUS, FLATIRON, zymgen, HealthTap, METABIOTA, ZEPHYR HEALTH, ovia, Ginger.io, transcriptic, Glow, enlitic, AiCure, Atomwise.
- Industries:** OPOWER, eHarmony, RetailNext, Compare in Store Analytics, duoetto, STITCH FIX, WorkFusion, TACHYUS, Seeq, FarmLogs, SwiftKey, HowGood, select, SIGHT MACHINE, statmuse, BOXEVER.

Footer: amazon web services, Google, Microsoft, IBM, SAP, sas, hp, Autonomy, vmware, talend, TIBC, TERADATA, ORACLE, NetApp.

```

public class LookupReducer extends Reducer<TextPair,Text,TextPair>
{
    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, Context context)
        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 corresponding values:
            else if(key.getSecond().toString().equals(Attribute.A1))
            {
                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() + ";" +
            attribute + ";" + product + ";" + trans_dt);

            if(start_dt_long < trans_dt_long && trans_dt_long < end_dt_long)
            {
                //concat output columns
                result = attribute + ";" + product + ";" + trans_dt;

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

    private static long dateToLong(String date){
        DateFormat formatter = new SimpleDateFormat("yyyy-MM-dd");
        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 WritableComparable<TextPair>
    {
        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
        public String toString(){
            return first.toString() + ";" + second.toString();
        }
    }
}

```

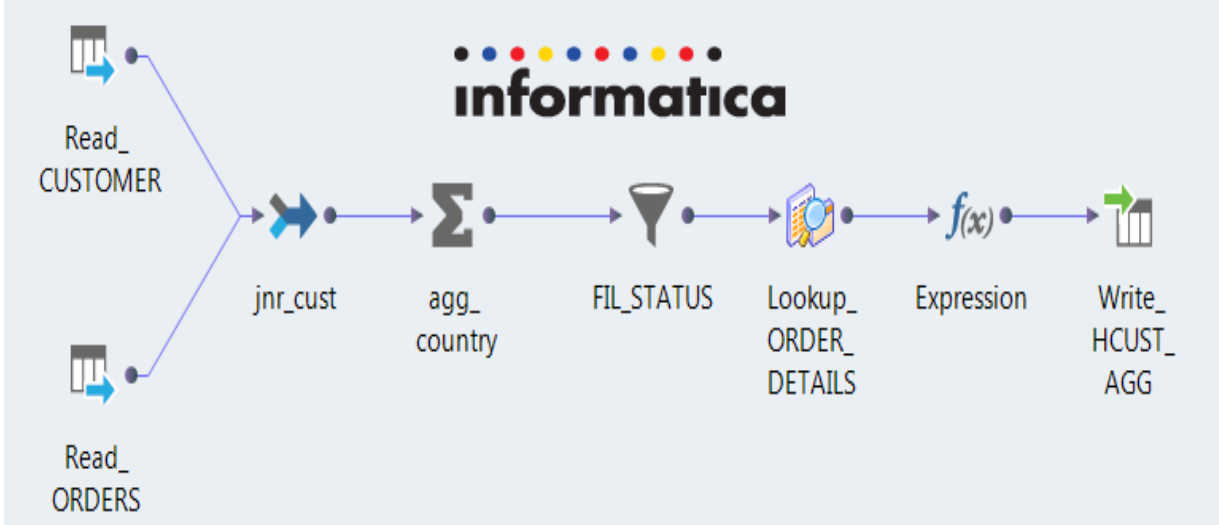
Универсальная платформа управления данными

Data Connectivity

Data Integration

Data Quality

Data Mastering

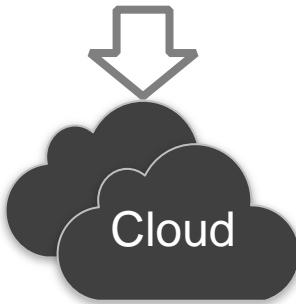
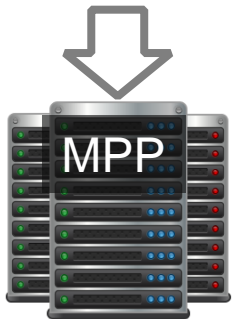
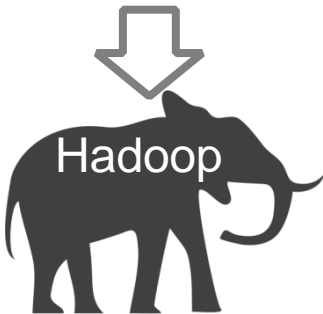
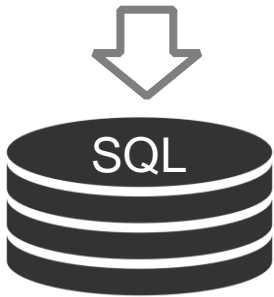


Self Service

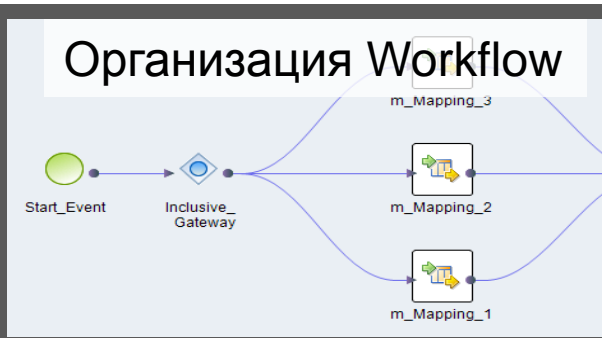
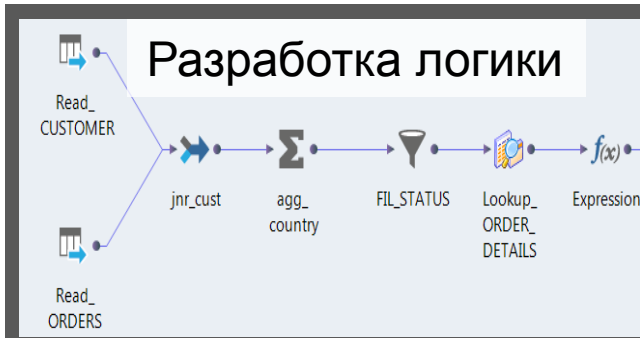
Data Governance

Data Security

Data Preparation



Использование Polyglot Engine



Мониторинг работы

A screenshot of a job monitoring interface. It shows a list of jobs with columns for 'Script', 'Status', 'User', and 'Time'. Below the list, there is a detailed view for a specific job with the ID 'infa_20130312084646_e4286942-c34a-43a6-af6c-bb0bc316c8d2'. The details include 'Started By: Administrator', 'User: Security Domain', 'Start Time: 03/12/2013 16:46:30', 'Elapsed Time: 00:00:15', and 'End Time: 03/12/2013 16:46:45'. At the bottom, there is a table for 'MR Job Details'.

Job ID	Map % Complete	Reduce % Complete
job_201303121452_0011	100	100

Smart Executor

Native

A box representing the Native Informatica Data Transformation Engine.

Кластер Hadoop

A diagram of a Hadoop cluster architecture. At the base are 'YARN' and 'HDFS'. Above them are four processing engines: 'Hive on Map Reduce' (with 'Map Reduce' below it), 'Hive On Tez' (with 'Tez' below it), 'Spark' (with 'Spark Core' below it), and 'Blaze' (with 'Cluster Aware' below it).

SQL

A box representing Database Pushdown.

Self Service для пользователей



informatica Live Data Map

Search Results

Filter by

- Asset Type
 - All
 - Column (57)
 - Column (41)
 - Data Domain (31)
 - Table (22)
- Resource Type
 - All
 - SAR Business Class...

Search Results (1 - 20 of 308)

- ORDER_DATE (Asset Type: Column)
- ORDER_DATE (Asset Type: Column)
- Orders (Asset Type: Class)
- ORDER_DATE

Last Updated: Dec 03, 2015 11:09am

Columns (4)

Name	Null Unique Non-Unique %	Source Data Type	Inferred Data Type	Data Domains
1 BRAND	0 11.77 88.23	VARCHAR2 (255)	String(20) 100.00%	AlphaNumeric_Spe... Date_AllForms I...
2 DESCRIPTION	0 49.22 7.78	VARCHAR2 (255)	String(89) 100.00%	Date_AllForms I... IPAddress N/A
3 PRICE	58.88 13.33 28.79	NUMBER (28)	Decimal(4) 100.00%	+2 more
4 PROD_ID	0 100 0	NUMBER (28)	Daily(yyddd) 47.77%	AlphaNumeric_Spe... +5 more

Workflow diagram showing a NewSQL source connected to a NewTarget target.

Lookup

Normalizer

Mapper

Aggregator

SQL

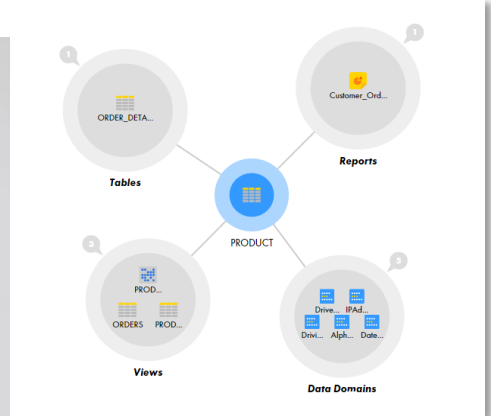
NewSQL Properties

General

Connection: MYDATABASE

Incoming Fields

SQL Type: Stored Procedure



Sample World Demo

id	address	city	state	zip	phone1	email	twitter	linkedin
1	8 W Carnton Ave #54	Bridgeport	NY	8054	856-636-8749	ant@barnes.org	@barnesdrucker	A Gold
2	189 Main St	Anchorage	AK	99501	907-386-4412	bsprague@format.com	@TheGOLTerre	B Silver
3	14 Carver St	Aurandale	OH	45011	60321-123-50-2819	shirley.alex@com.net	@shirleyalexand	C Bronze
4	3 Kincaid Dr	Ashland	OH	44805	44805-439-563-2884	emona@emona.com	@Ematrick	B Silver
5	7 Eads St	Chicago	IL	60612	60612-775-575-8514	mlisa_culhan@yahoo.com	@mlisculhan	C Bronze
6	716 Jackson Blvd	San Jose	CA	95111	95111-408-762-3500	john@john.com	@johnjohn	C Bronze
7	5 Sutton Ave #88	Scotts Falls	SD	57335	57335-805-416-2147	wage_meyer@com.net	@meyerwage	B Silver
8	228 Runnuck Pl #200	Baltimore	MD	21224	21224-438-658-8723	lisa@lisa.com	@LISAPEACHES	C Bronze
9	2371 Arnold Ave	Beltsville	PA	18463	18463-215-674-3205	norma_prigon@yahoo.com	@PrigonNorma	A Gold
10	12725 St 10 21st M	Middle Island	NY	11953	11953-635-335-5614	awanda@l@gmail.com	@awandaawanda	C Bronze
11	251 E 75th St #89	San Jose	CA	95034	95034-338-688-5611	lucy.culhan@att.com	@LucyCulhan29	A Gold
12	391 388 Conventicut Ave Ne	Chagrin Falls	OH	44023	44023-448-780-4425	graham@com.net	@Graham29177	A Gold
13	361 Milwaukee St	San Jose	TX	79045	79045-956-517-6135	colleen@digital.com	@colleencolleen	B Silver
14	75 State Road 434 E	Phoenix	AZ	85013	85013-602-277-4380	monty@ltd.com	@rightmonty	B Silver
15	48974 E Carrillo St	San Minerva	TN	37110	37110-915-913-9635	maggie@format.com	@maggie29	B Silver
16	123 New Avenue Blvd	Minneapolis	WI	53057	53057-414-661-6508	andrea_walsh@com.net	@andreaandrea	A Gold

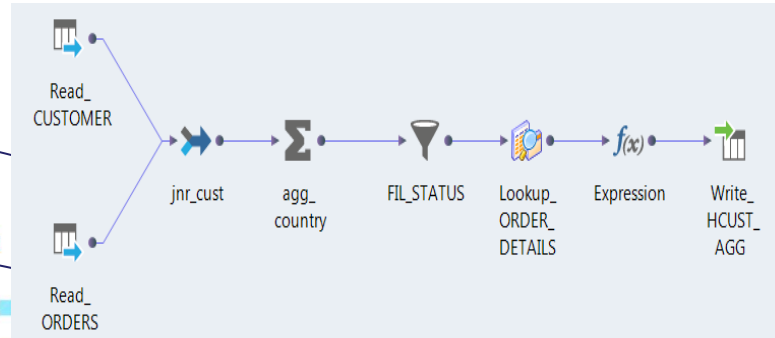
Column Overview

Column	Type	US City #
City	CHAR(17)	8
State	CHAR(2)	0
Zip	CHAR(5)	7
Length Range	4 to 19	0

Value Frequencies

Value	Frequency
New York	14
Philadelphia	8
Chicago	7
Miami	6
San Francisco	5
Carolina	5
Orlando	5
Phoenix	5

Потоковая онлайн-аналитика



Kafka

Alerting

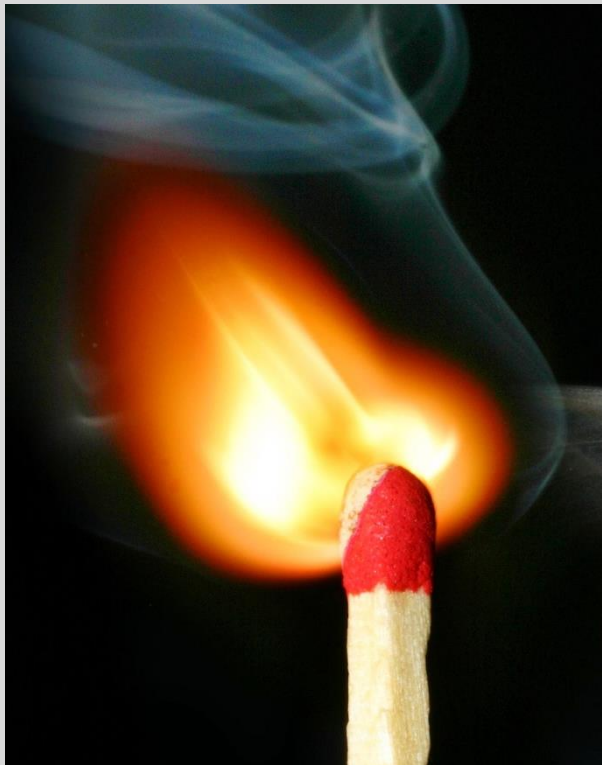
Transformations

Rules Engine

Spark Streaming

Kafka





#1 Time-To-Market

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

#2 Надежность

- Снижение рисков
- Полный контроль
- Поддержка
- Возможность развития







Спасибо за внимание!
Ваши вопросы

DIS Group

Петр Борисов

info@dis-group.ru

+7 495 645 02 01