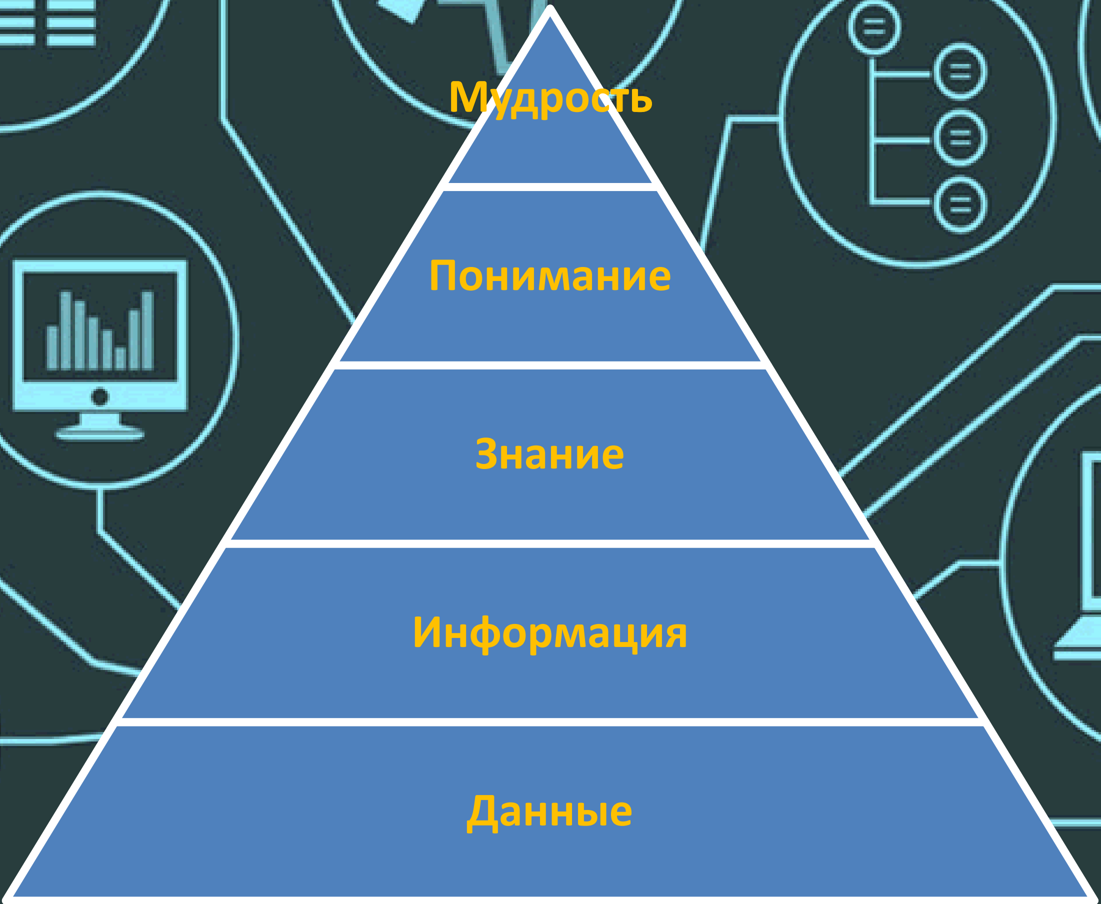
Методы работы с большими данными

Киреев Василий Сергеевич,

к.т.н., доцент

Москва, 2020

Пирамида знаний



Мудрость

Понимание

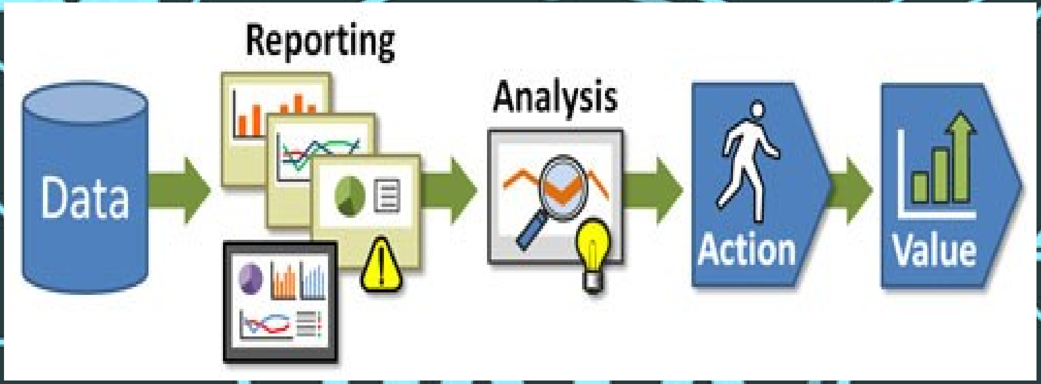
Знание

Информация

Данные

DATA-driven

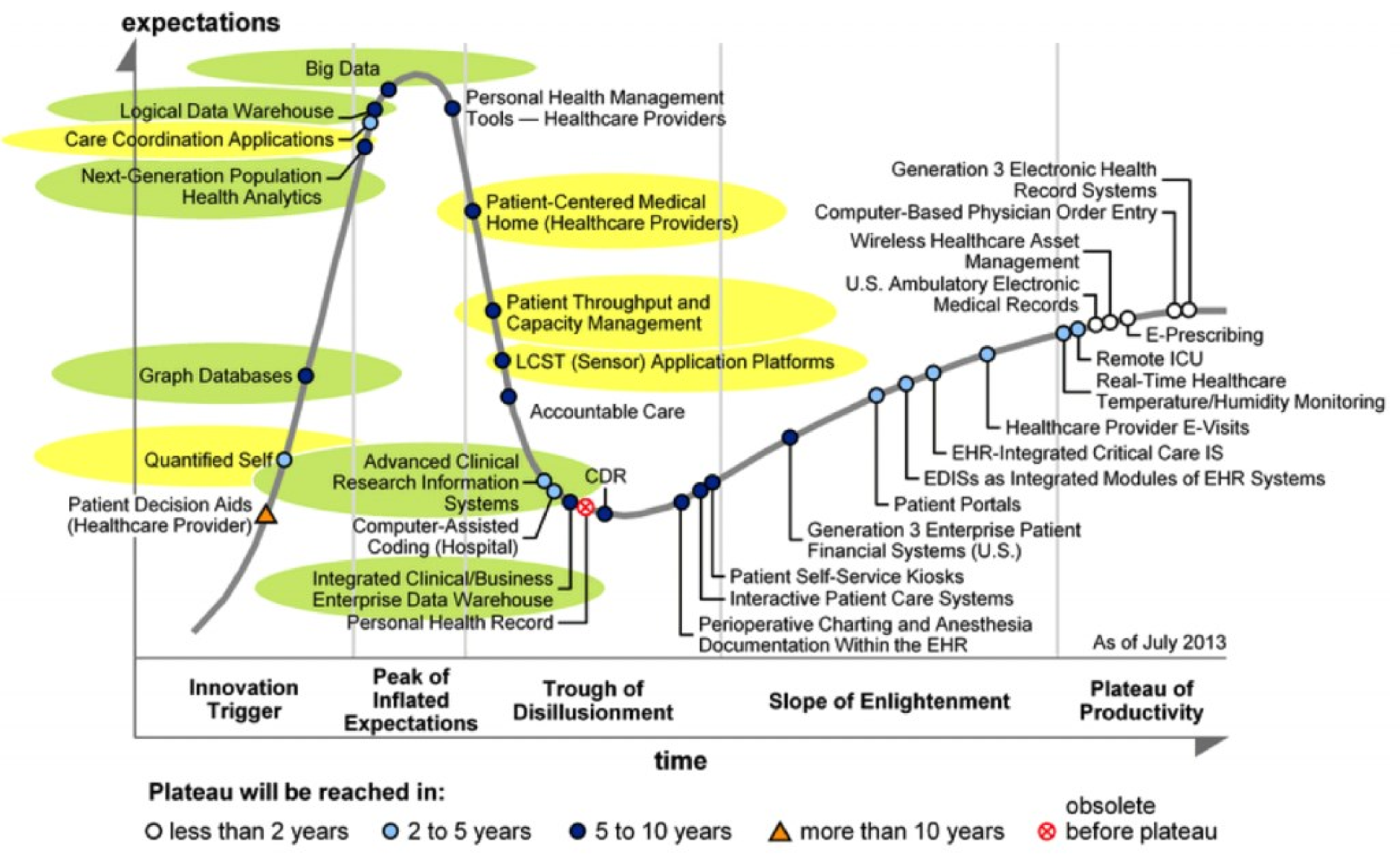
- **Data Driven (дословно — «управляемый данными»)** — это подход к управлению, основанный на данных. Его главный постулат: решения нужно принимать, опираясь на анализ цифр, а не интуицию и личный опыт.



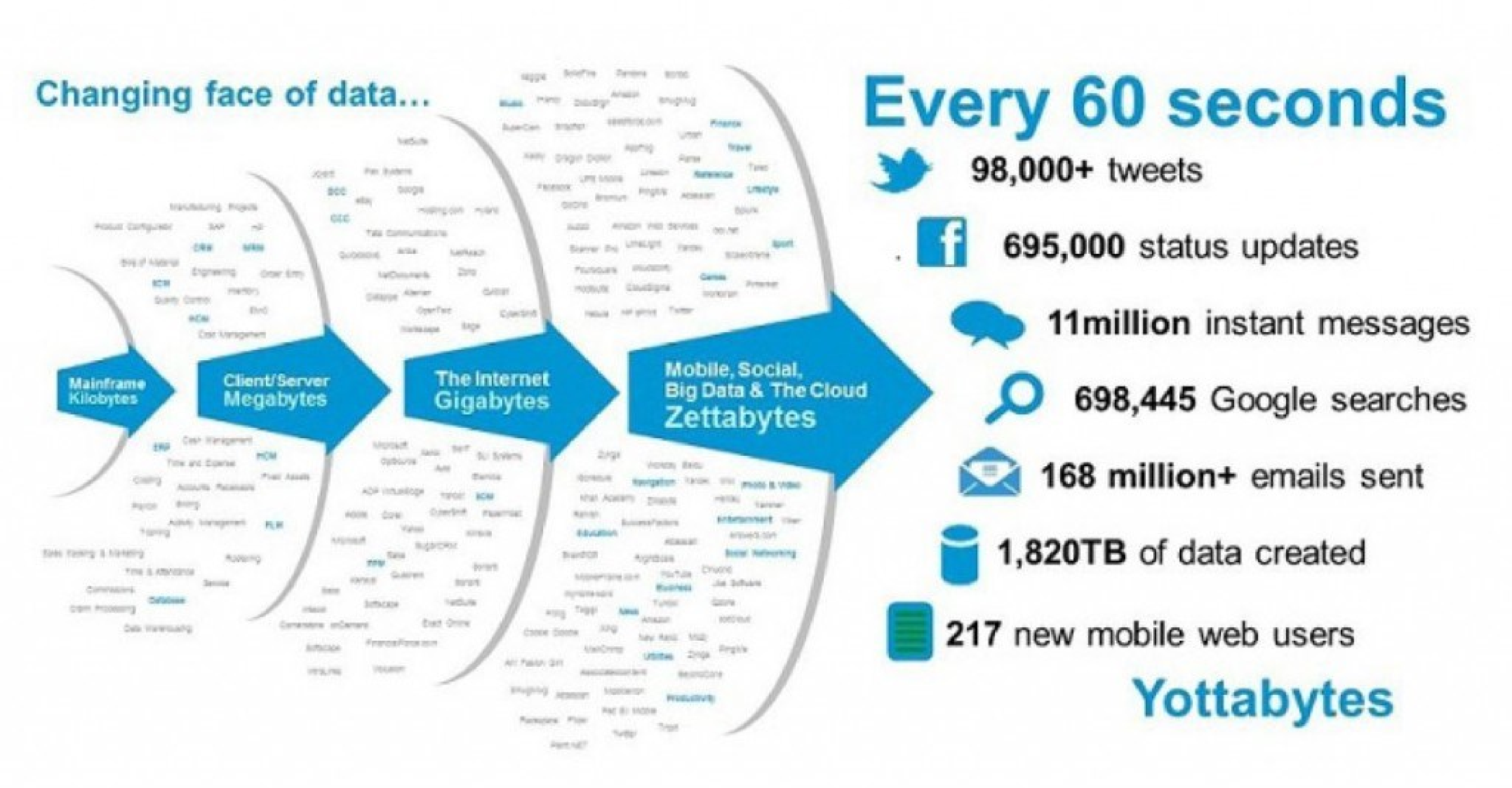
Большие Данные

- Big Data is a term used to describe the large amount of data in the networked, digitized, sensor-laden, information-driven world *(NIST)*
- Термин «Большие данные» относится к наборам данных, размер которых превосходит возможности типичных баз данных (БД) по занесению, хранению, управлению и анализу информации

Большие Данные



Что такое большие данные?



Changing face of data…

Every 60 seconds

98,000+ tweets

695,000 status updates

11million instant messages

698,445 Google searches

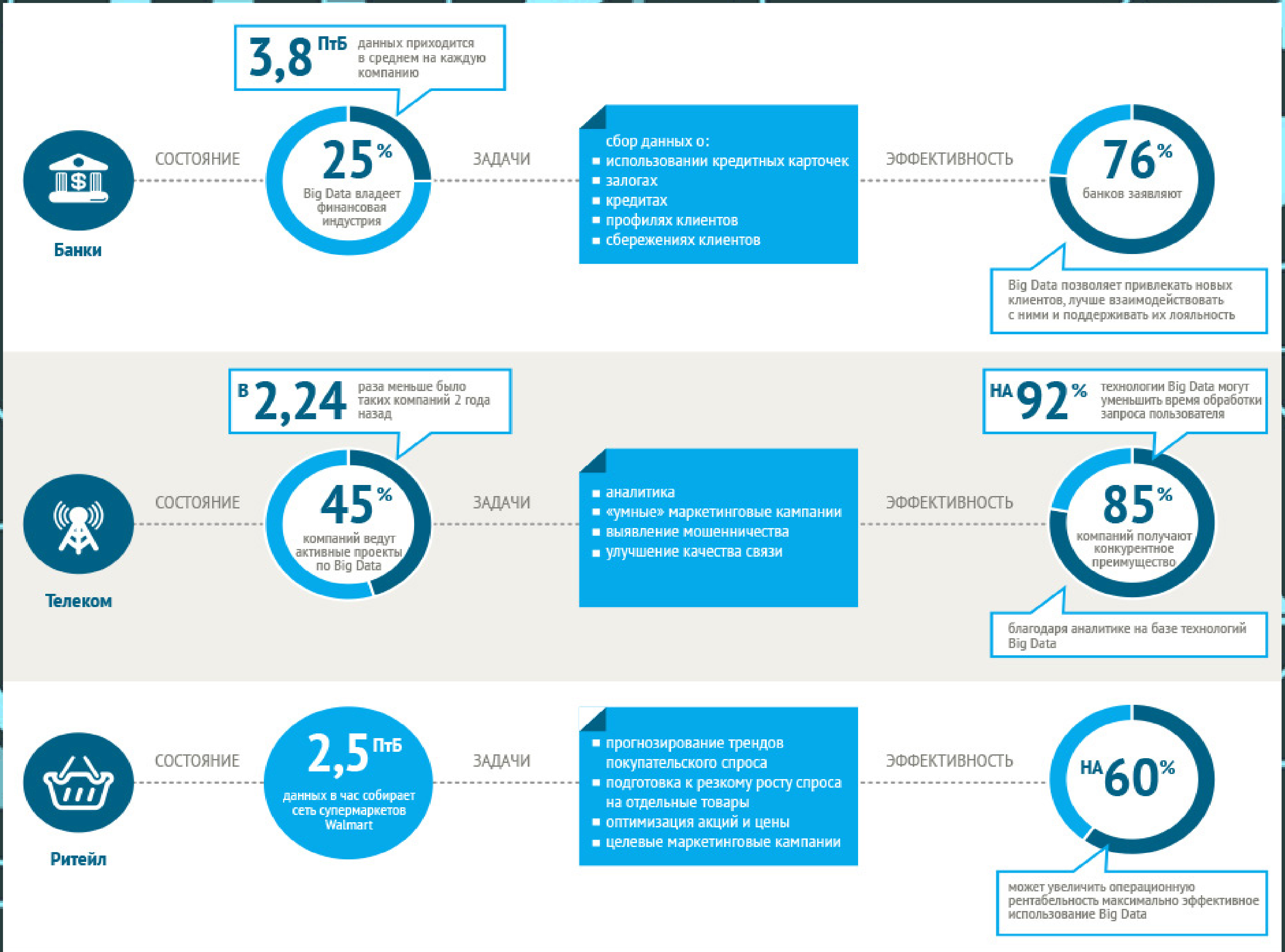
168 million+ emails sent

1,820TB of data created

217 new mobile web users

Yottabytes

Большие данные в отраслях



Банки

СОСТОЯНИЕ

3,8 ПтБ данных приходится в среднем на каждую компанию

25% Big Data владеет финансовая индустрия

ЗАДАЧИ

сбор данных о:
- использовании кредитных карточек
- залогах
- кредитах
- профилях клиентов
- сбережениях клиентов

ЭФФЕКТИВНОСТЬ

76% банков заявляют

Big Data позволяет привлекать новых клиентов, лучше взаимодействовать с ними и поддерживать их лояльность

Телеком

СОСТОЯНИЕ

В 2,24 раза меньше было таких компаний 2 года назад

45% компаний ведут активные проекты по Big Data

ЗАДАЧИ

- аналитика
- «умные» маркетинговые кампании
- выявление мошенничества
- улучшение качества связи

ЭФФЕКТИВНОСТЬ

НА 92% технологии Big Data могут уменьшить время обработки запроса пользователя

85% компаний получают конкурентное преимущество

благодаря аналитике на базе технологий Big Data

Ритейл

СОСТОЯНИЕ

2,5 ПтБ данных в час собирает сеть супермаркетов Walmart

ЗАДАЧИ

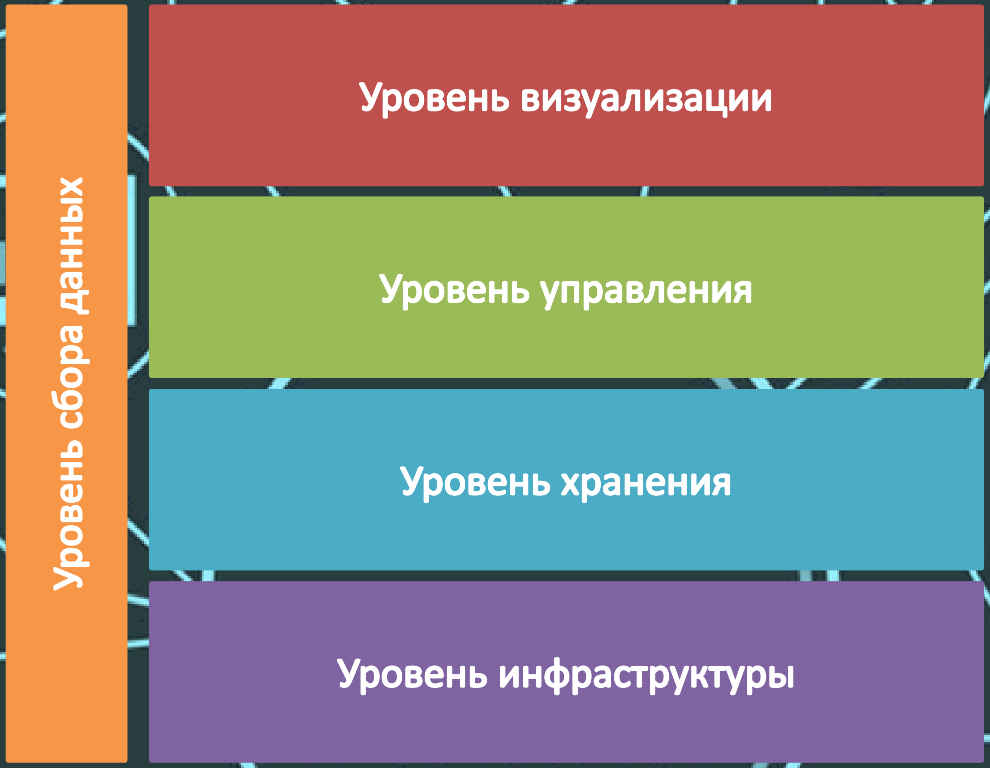
- прогнозирование трендов покупательского спроса
- подготовка к резкому росту спроса на отдельные товары
- оптимизация акций и цены
- целевые маркетинговые кампании

ЭФФЕКТИВНОСТЬ

НА 60%

может увеличить операционную рентабельность максимально эффективное использование Big Data

Уровни архитектуры по обращению с Большими данными



Уровень сбора данных

Уровень визуализации

Уровень управления

Уровень хранения

Уровень инфраструктуры

Уровень сбора данных

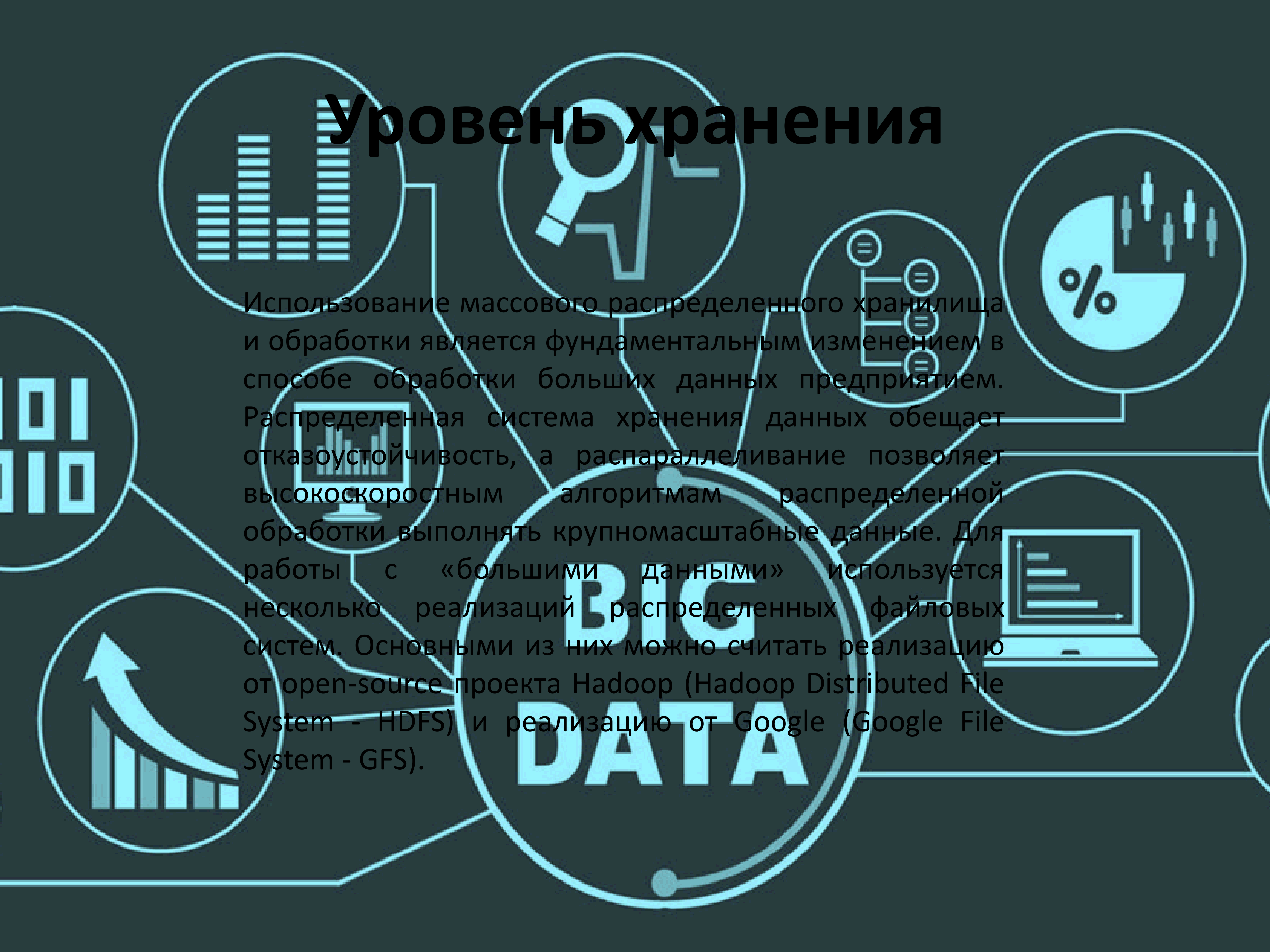
Этот уровень отвечает за отделение шума от соответствующей информации, а также регулирование объема, скорости и разнообразия данных. Он должен иметь возможность проверять, очищать, преобразовывать, уменьшать и интегрировать данные в стек технологий больших данных для дальнейшей обработки. Это новое программное обеспечение, которое должно быть масштабируемым, устойчивым, отзывчивым и регулирующим в архитектуре больших данных.

Уровень инфраструктуры

На данном уровне располагается физическая инфраструктура, необходимая для функционирования и масштабируемости архитектуры больших данных. Фактически наличие надежной и недорогой физической инфраструктуры привело к появлению таких важных тенденций, как big data. Для поддержки непредвиденного или непредсказуемого объема, скорости или разнообразия данных физическая инфраструктура для больших данных должна отличаться от инфраструктуры для традиционных данных.

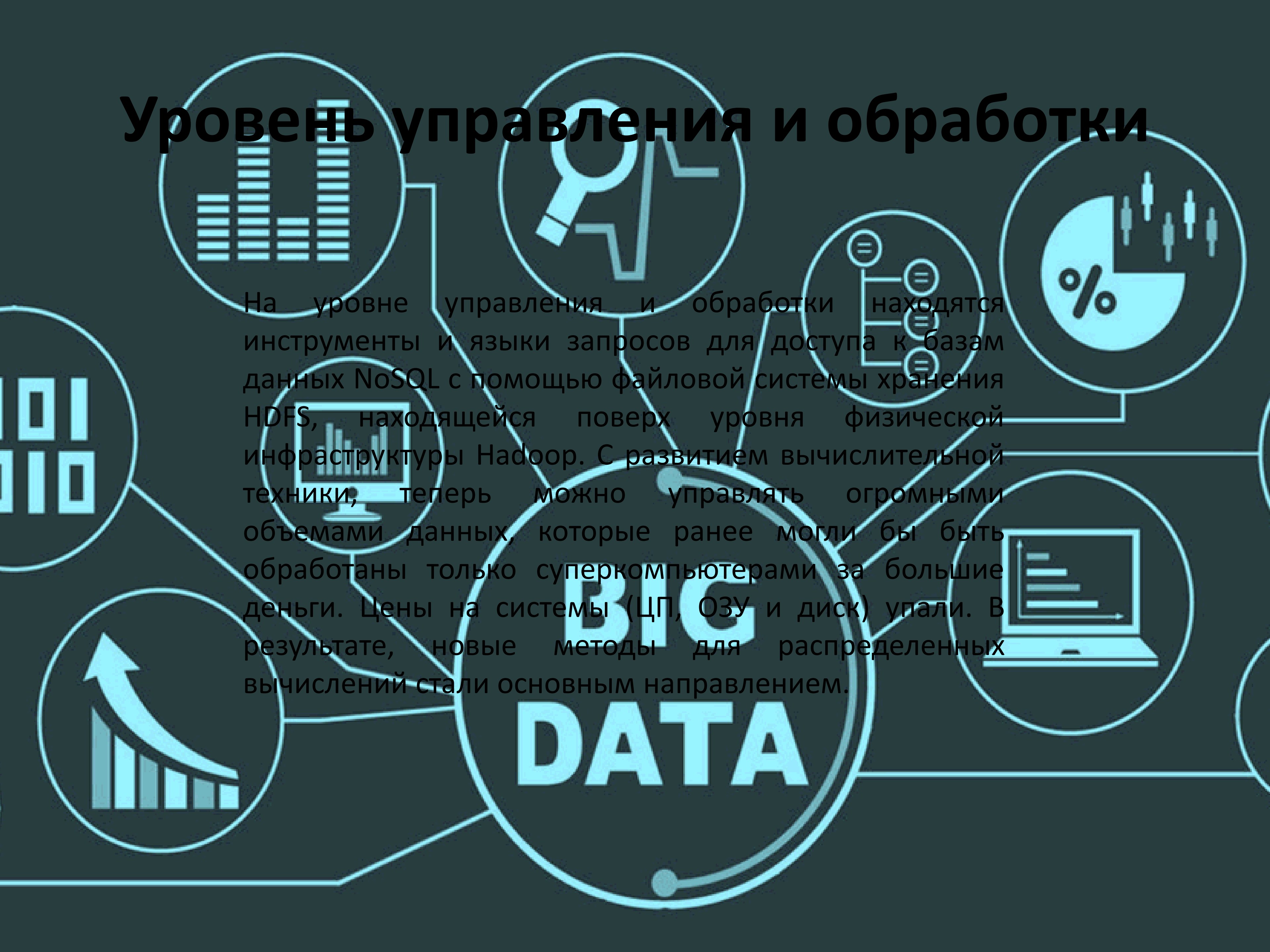
Уровень хранения

Использование массового распределенного хранилища и обработки является фундаментальным изменением в способе обработки больших данных предприятием. Распределенная система хранения данных обещает отказоустойчивость, а распараллеливание позволяет высокоскоростным алгоритмам распределенной обработки выполнять крупномасштабные данные. Для работы с «большими данными» используется несколько реализаций распределенных файловых систем. Основными из них можно считать реализацию от open-source проекта Hadoop (Hadoop Distributed File System - HDFS) и реализацию от Google (Google File System - GFS).

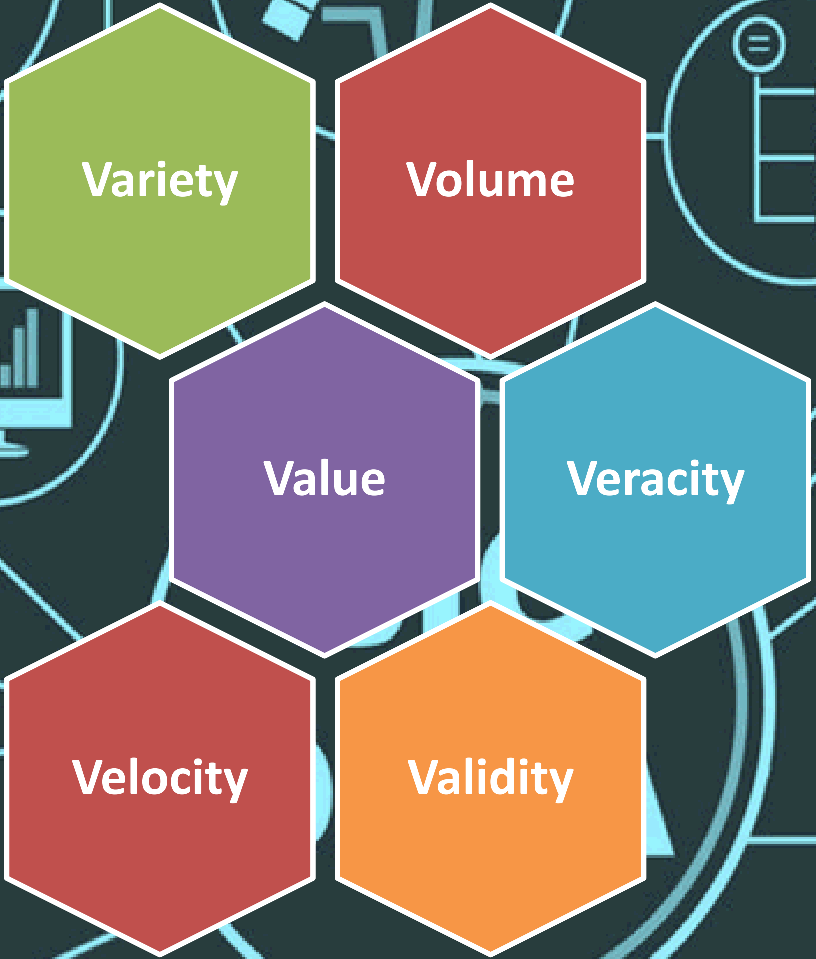


Уровень управления и обработки

На уровне управления и обработки находятся инструменты и языки запросов для доступа к базам данных NoSQL с помощью файловой системы хранения HDFS, находящейся поверх уровня физической инфраструктуры Hadoop. С развитием вычислительной техники, теперь можно управлять огромными объемами данных, которые ранее могли бы быть обработаны только суперкомпьютерами за большие деньги. Цены на системы (ЦП, ОЗУ и диск) упали. В результате, новые методы для распределенных вычислений стали основным направлением.



V-модель Больших Данных



Volume

- Гигабайты – 1024 мегабайт
- Терабайты – 1024 гигабайт
- Петабайты – 1024 терабайт
- Эксобайты – 1024 петабайт
- Зетабайты – 1024 эксобайт

Variety

- Тексты
- Изображения
- Видео
- Аудио
- Показания датчиков
- Транзакции
- Гео-данные

Velocity

- Пакетная обработка
- Стриминговая обработка
- Обработка в реальном времени

Veracity

- Отсутствие целостности
- Отсутствие полноты
- Отсутствие определенности

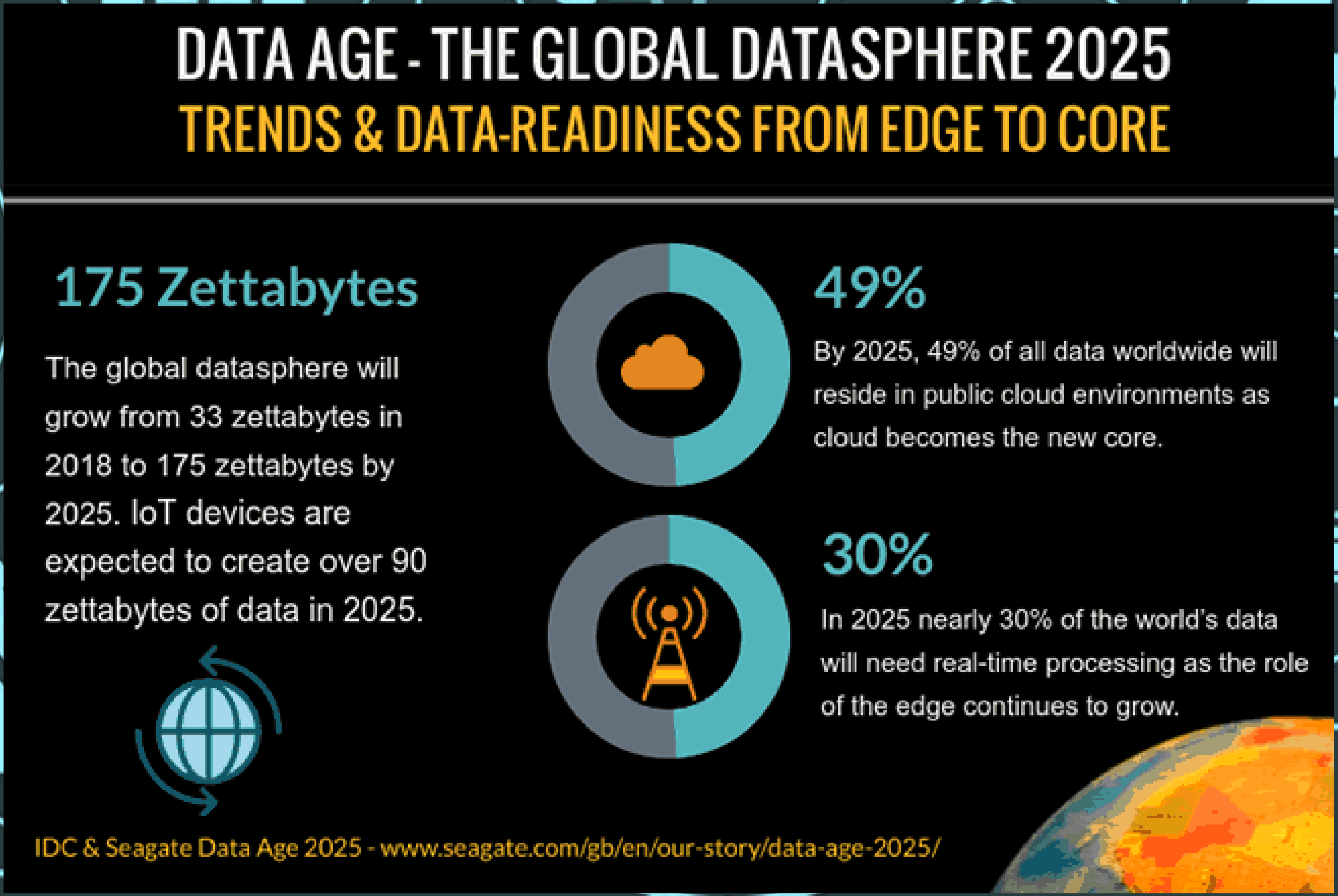
Value

- Статистическая ценность
- Финансовая ценность
- Ценность для продукта

Validity

- Прозрачность
- Законность
- Конфиденциальность

Тренды в Больших Данных



DATA AGE - THE GLOBAL DATASPHERE 2025

TRENDS & DATA-READINESS FROM EDGE TO CORE

175 Zettabytes

The global datasphere will grow from 33 zettabytes in 2018 to 175 zettabytes by 2025. IoT devices are expected to create over 90 zettabytes of data in 2025.

49%

By 2025, 49% of all data worldwide will reside in public cloud environments as cloud becomes the new core.

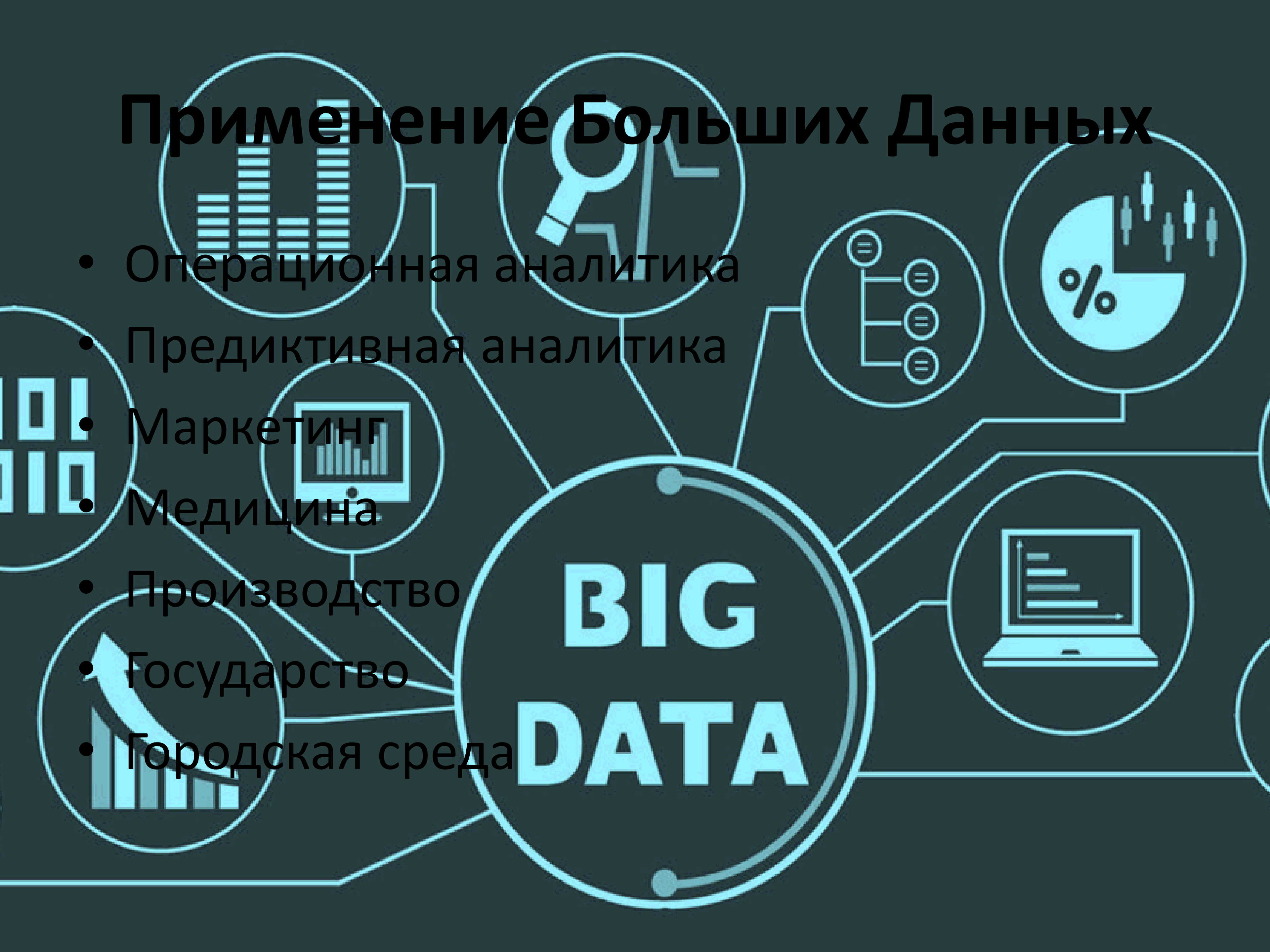
30%

In 2025 nearly 30% of the world's data will need real-time processing as the role of the edge continues to grow.

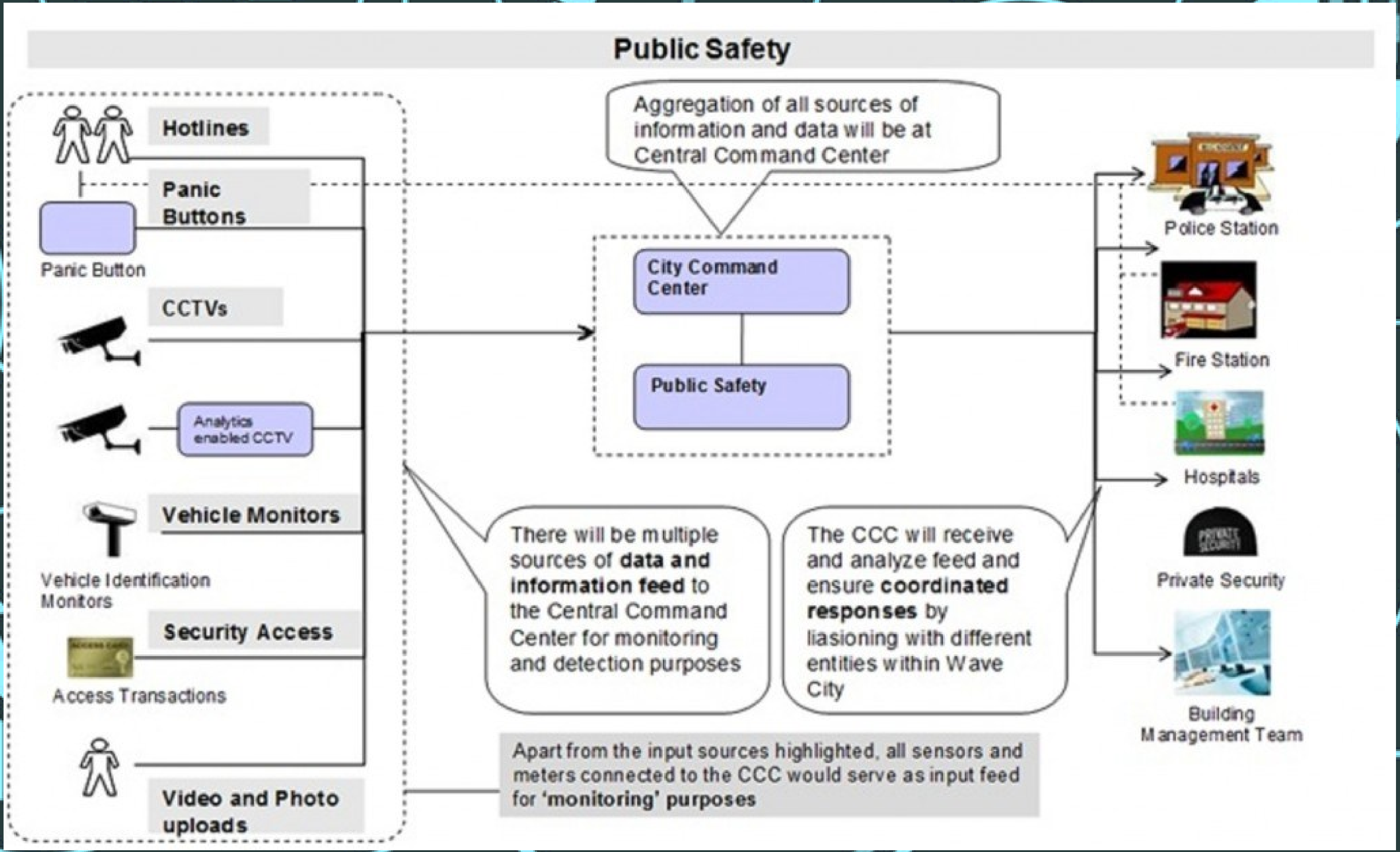
IDC & Seagate Data Age 2025 - www.seagate.com/gb/en/our-story/data-age-2025/

Применение Больших Данных

- Операционная аналитика
- Предиктивная аналитика
- Маркетинг
- Медицина
- Производство
- Государство
- Городская среда



Общественная безопасность



Public Safety

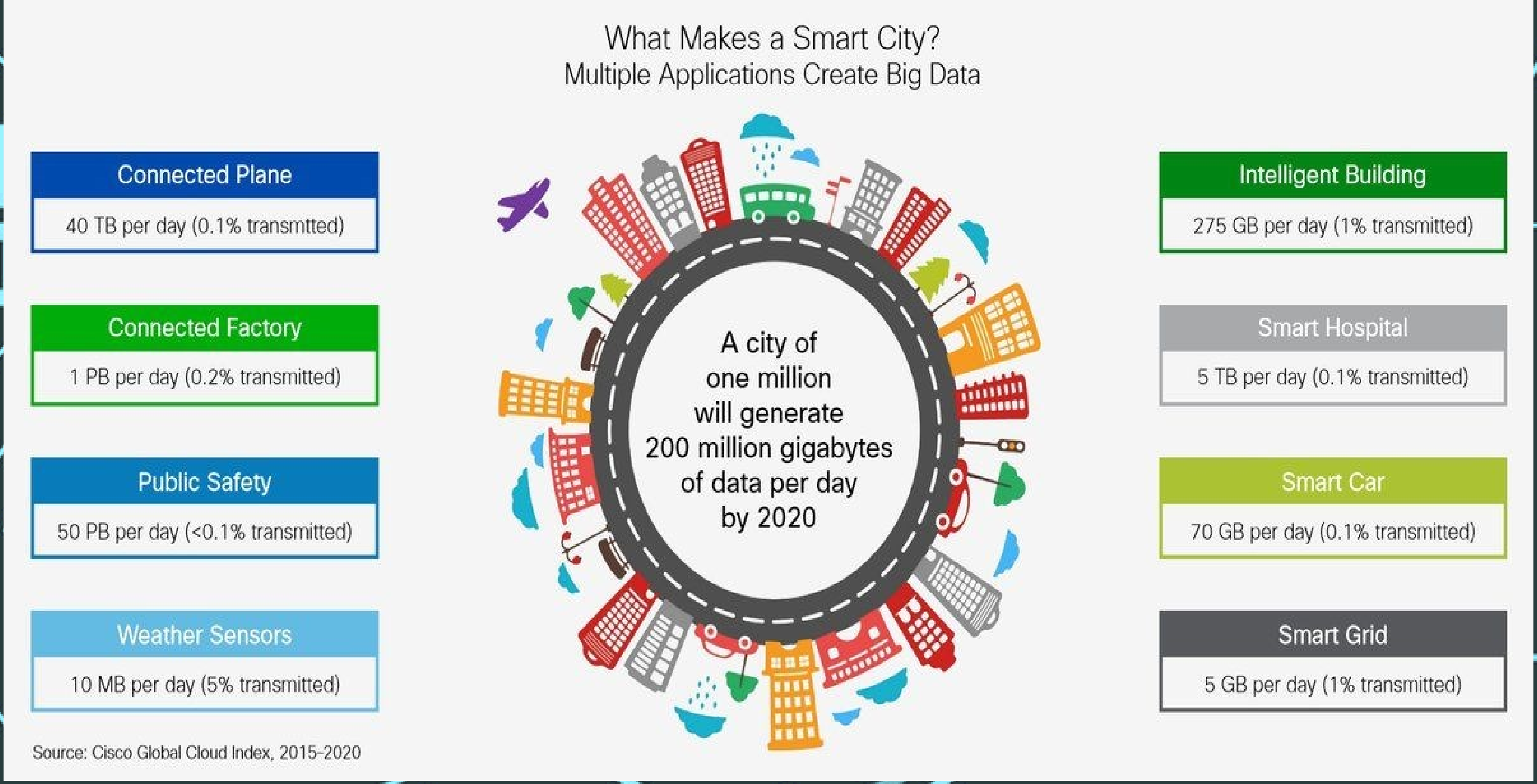
Aggregation of all sources of information and data will be at Central Command Center

There will be multiple sources of **data and information feed** to the Central Command Center for monitoring and detection purposes

The CCC will receive and analyze feed and ensure **coordinated responses** by liasioning with different entities within Wave City

Apart from the input sources highlighted, all sensors and meters connected to the CCC would serve as input feed for **'monitoring' purposes**

Транспорт и городская инфраструктура



What Makes a Smart City?
Multiple Applications Create Big Data

Connected Plane
40 TB per day (0.1% transmtted)

Connected Factory
1 PB per day (0.2% transmitted)

Public Safety
50 PB per day (<0.1% transmitted)

Weather Sensors
10 MB per day (5% transmitted)

A city of one million will generate 200 million gigabytes of data per day by 2020

Intelligent Building
275 GB per day (1% transmitted)

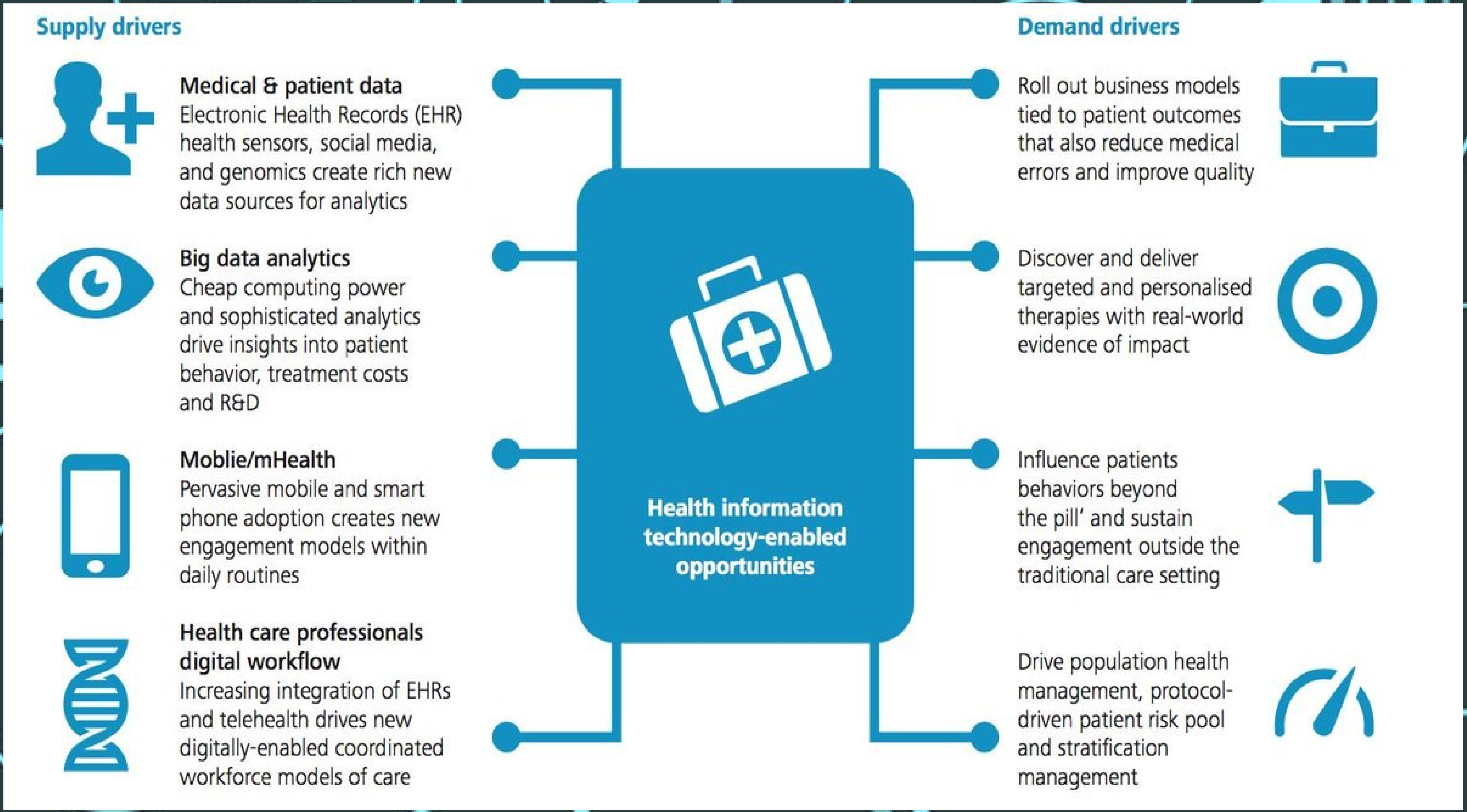
Smart Hospital
5 TB per day (0.1% transmitted)

Smart Car
70 GB per day (0.1% transmitted)

Smart Grid
5 GB per day (1% transmitted)

Source: Cisco Global Cloud Index, 2015-2020

Общественное здравоохранение



Supply drivers

Medical & patient data
Electronic Health Records (EHR) health sensors, social media, and genomics create rich new data sources for analytics

Big data analytics
Cheap computing power and sophisticated analytics drive insights into patient behavior, treatment costs and R&D

Moblie/mHealth
Pervasive mobile and smart phone adoption creates new engagement models within daily routines

Health care professionals digital workflow
Increasing integration of EHRs and telehealth drives new digitally-enabled coordinated workforce models of care

Health information technology-enabled opportunities

Demand drivers

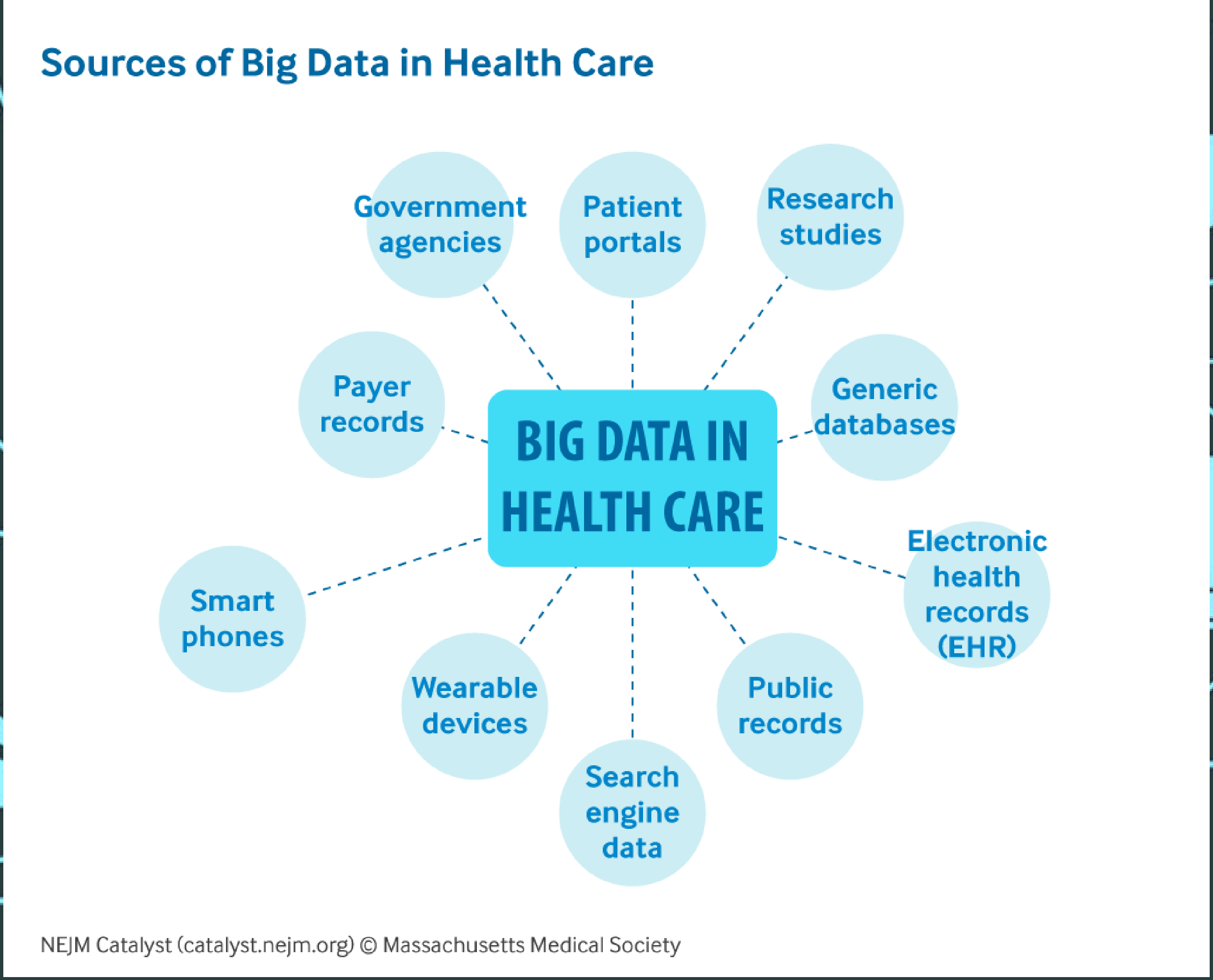
Roll out business models tied to patient outcomes that also reduce medical errors and improve quality

Discover and deliver targeted and personalised therapies with real-world evidence of impact

Influence patients behaviors beyond the pill' and sustain engagement outside the traditional care setting

Drive population health management, protocol-driven patient risk pool and stratification management

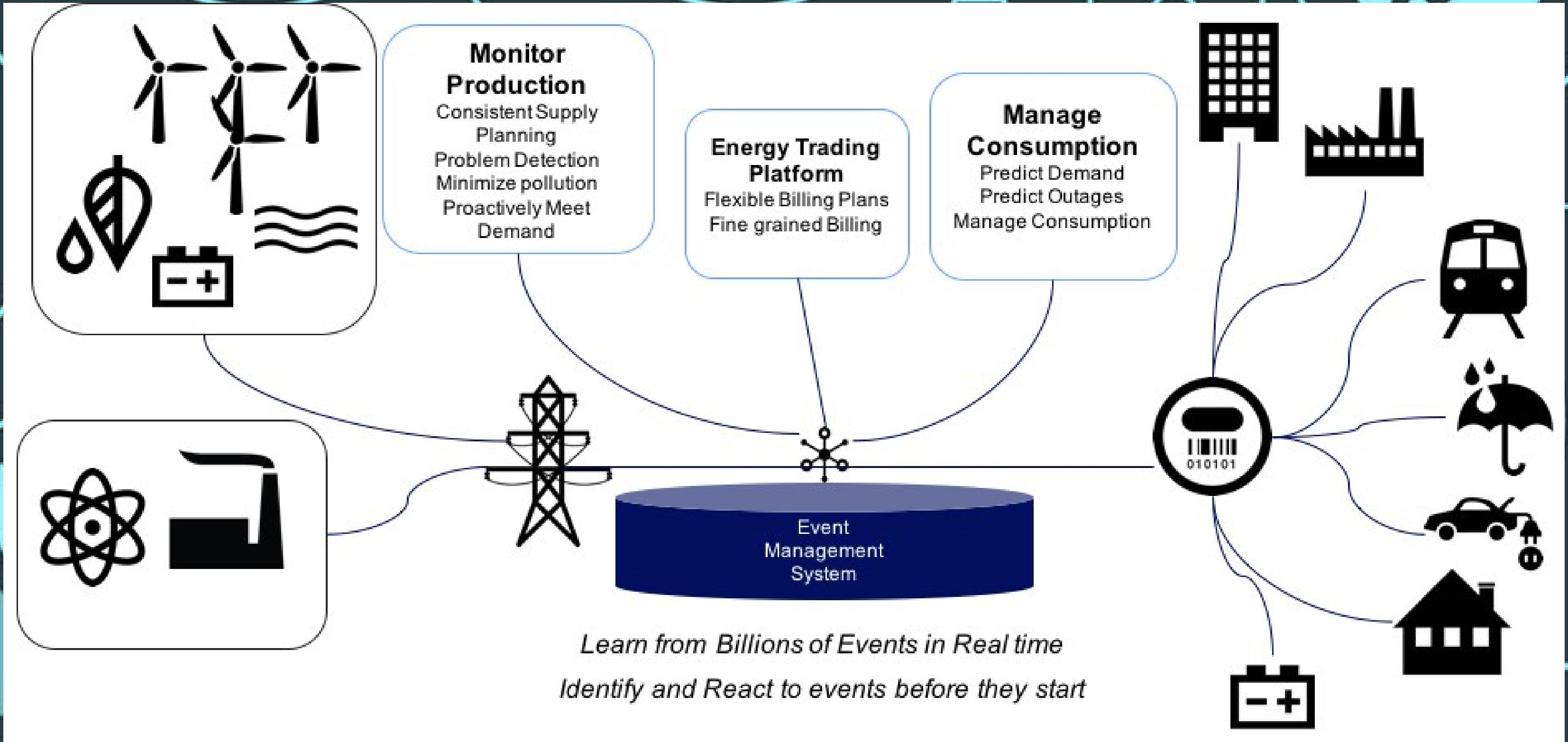
Общественное здравоохранение



Sources of Big Data in Health Care

NEJM Catalyst (catalyst.nejm.org) © Massachusetts Medical Society

Энергетика, окружающая среда и ЖКХ



Monitor Production
Consistent Supply Planning
Problem Detection
Minimize pollution
Proactively Meet Demand

Energy Trading Platform
Flexible Billing Plans
Fine grained Billing

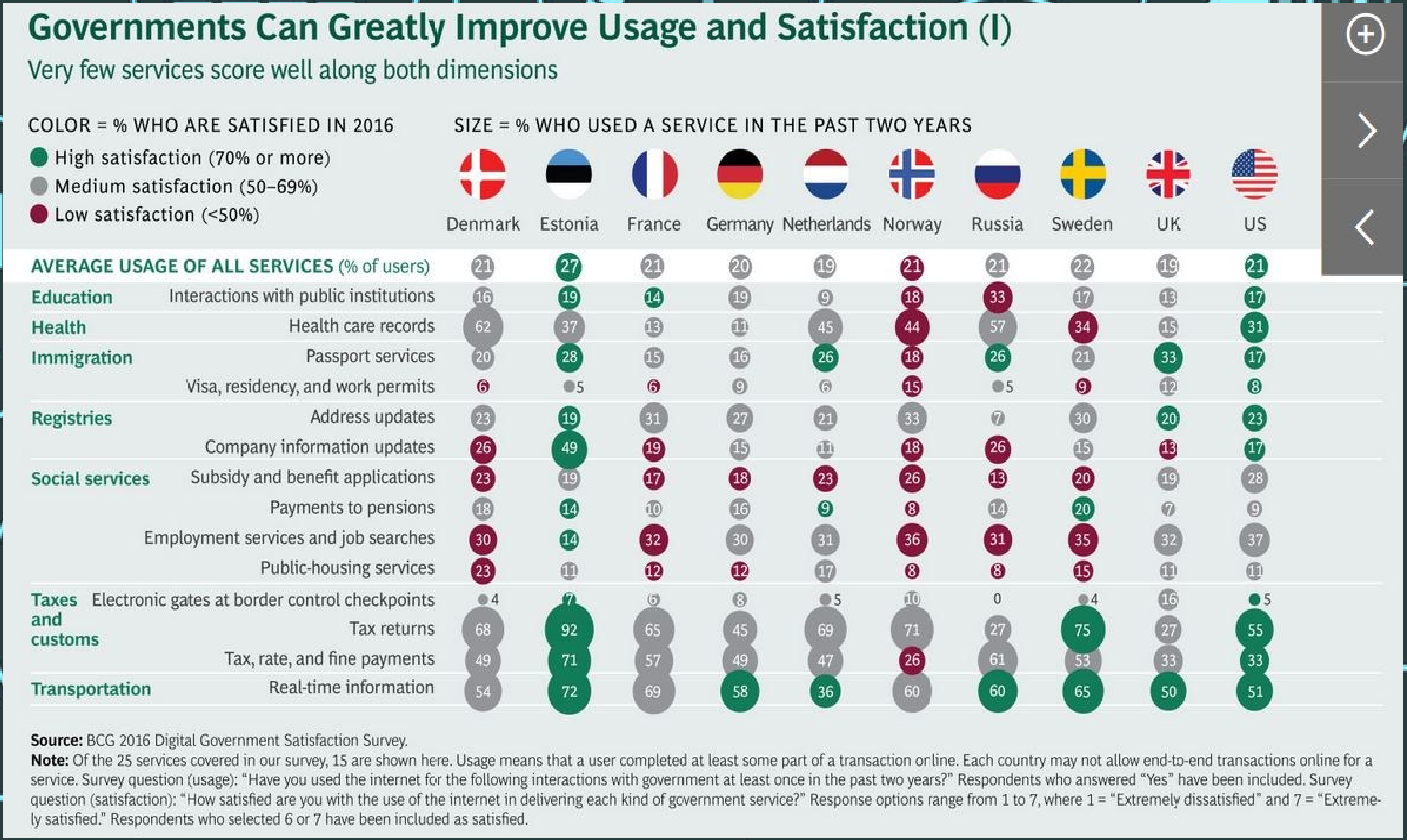
Manage Consumption
Predict Demand
Predict Outages
Manage Consumption

Event Management System

Learn from Billions of Events in Real time

Identify and React to events before they start

Цифровые государственные услуги в мире



Governments Can Greatly Improve Usage and Satisfaction (I)

Very few services score well along both dimensions

COLOR = % WHO ARE SATISFIED IN 2016

- 🟢 High satisfaction (70% or more)
- ⚪ Medium satisfaction (50–69%)
- 🔴 Low satisfaction (<50%)

SIZE = % WHO USED A SERVICE IN THE PAST TWO YEARS

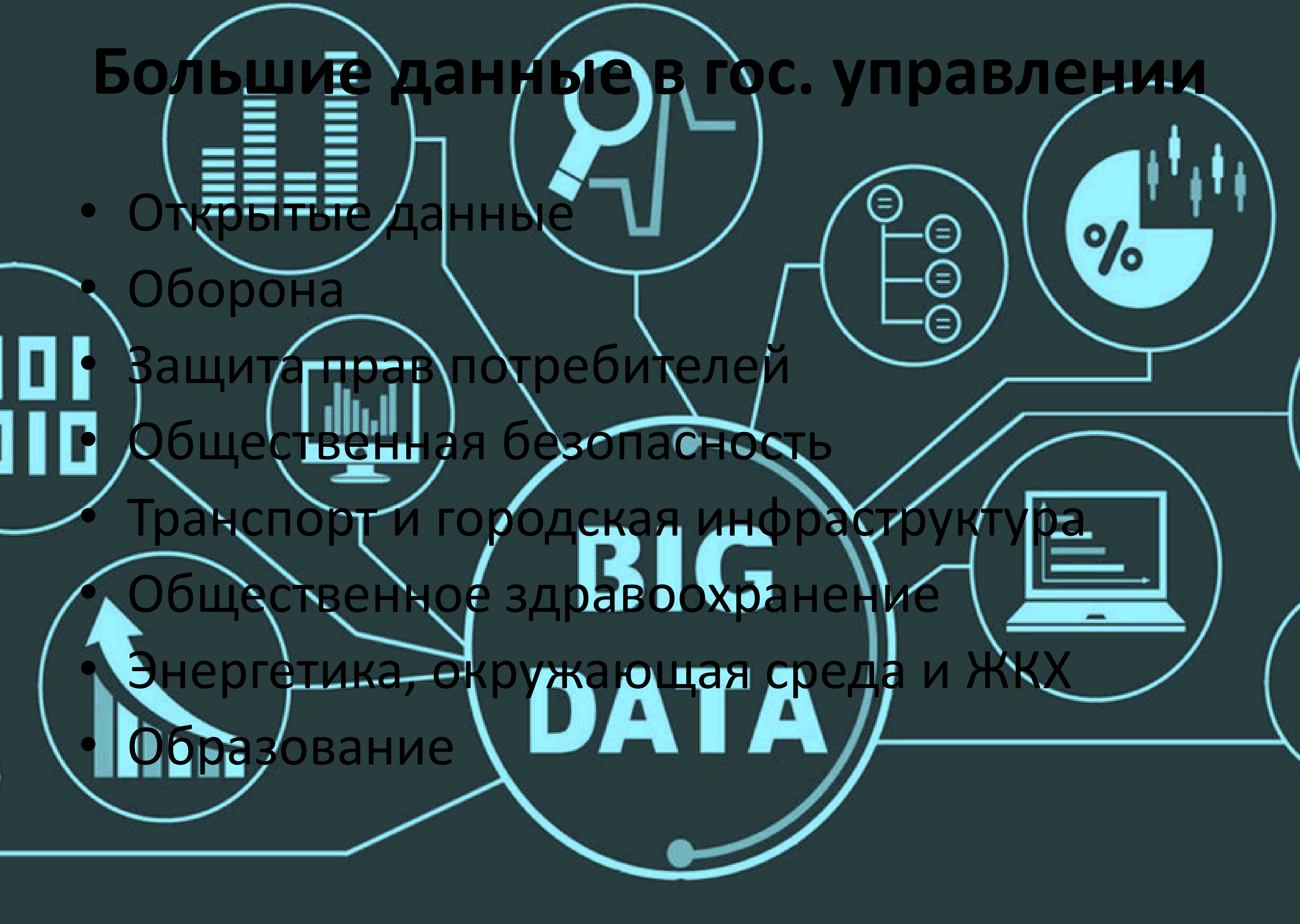
		Denmark	Estonia	France	Germany	Netherlands	Norway	Russia	Sweden	UK	US
AVERAGE USAGE OF ALL SERVICES (% of users)		⚪ 21	🟢 27	⚪ 21	⚪ 20	⚪ 19	🔴 21	⚪ 21	⚪ 22	⚪ 19	🟢 21
Education	Interactions with public institutions	⚪ 16	🟢 19	🟢 14	⚪ 19	⚪ 9	🔴 18	🔴 33	⚪ 17	⚪ 13	🟢 17
Health	Health care records	⚪ 62	⚪ 37	⚪ 13	⚪ 11	⚪ 45	🔴 44	⚪ 57	🔴 34	⚪ 15	🟢 31
Immigration	Passport services	⚪ 20	🟢 28	⚪ 15	⚪ 16	🟢 26	🔴 18	🟢 26	⚪ 21	🟢 33	🟢 17
	Visa, residency, and work permits	🔴 6	⚪ 5	🔴 6	⚪ 9	⚪ 6	🔴 15	⚪ 5	🔴 9	⚪ 12	🟢 8
Registries	Address updates	⚪ 23	🟢 19	⚪ 31	⚪ 27	⚪ 21	⚪ 33	⚪ 7	⚪ 30	🟢 20	🟢 23
	Company information updates	🔴 26	🟢 49	🔴 19	⚪ 15	⚪ 11	🔴 18	🔴 26	⚪ 15	🔴 13	🟢 17
Social services	Subsidy and benefit applications	🔴 23	⚪ 19	🔴 17	🔴 18	🔴 23	🔴 26	🔴 13	🔴 20	⚪ 19	⚪ 28
	Payments to pensions	⚪ 18	🟢 14	⚪ 10	⚪ 16	🟢 9	🔴 8	⚪ 14	🟢 20	⚪ 7	⚪ 9
	Employment services and job searches	🔴 30	🟢 14	🔴 32	⚪ 30	⚪ 31	🔴 36	🔴 31	🔴 35	⚪ 32	⚪ 37
	Public-housing services	🔴 23	⚪ 11	🔴 12	🔴 12	⚪ 17	🔴 8	🔴 8	🔴 15	⚪ 11	⚪ 11
Taxes and customs	Electronic gates at border control checkpoints	⚪ 4	🟢 7	⚪ 6	⚪ 8	⚪ 5	⚪ 10	0	⚪ 4	⚪ 16	🟢 5
	Tax returns	⚪ 68	🟢 92	⚪ 65	⚪ 45	⚪ 69	⚪ 71	⚪ 27	🟢 75	⚪ 27	🟢 55
	Tax, rate, and fine payments	⚪ 49	🟢 71	⚪ 57	⚪ 49	⚪ 47	🔴 26	⚪ 61	⚪ 53	⚪ 33	🟢 33
Transportation	Real-time information	⚪ 54	🟢 72	⚪ 69	🟢 58	🟢 36	⚪ 60	🟢 60	🟢 65	🟢 50	🟢 51

Source: BCG 2016 Digital Government Satisfaction Survey.

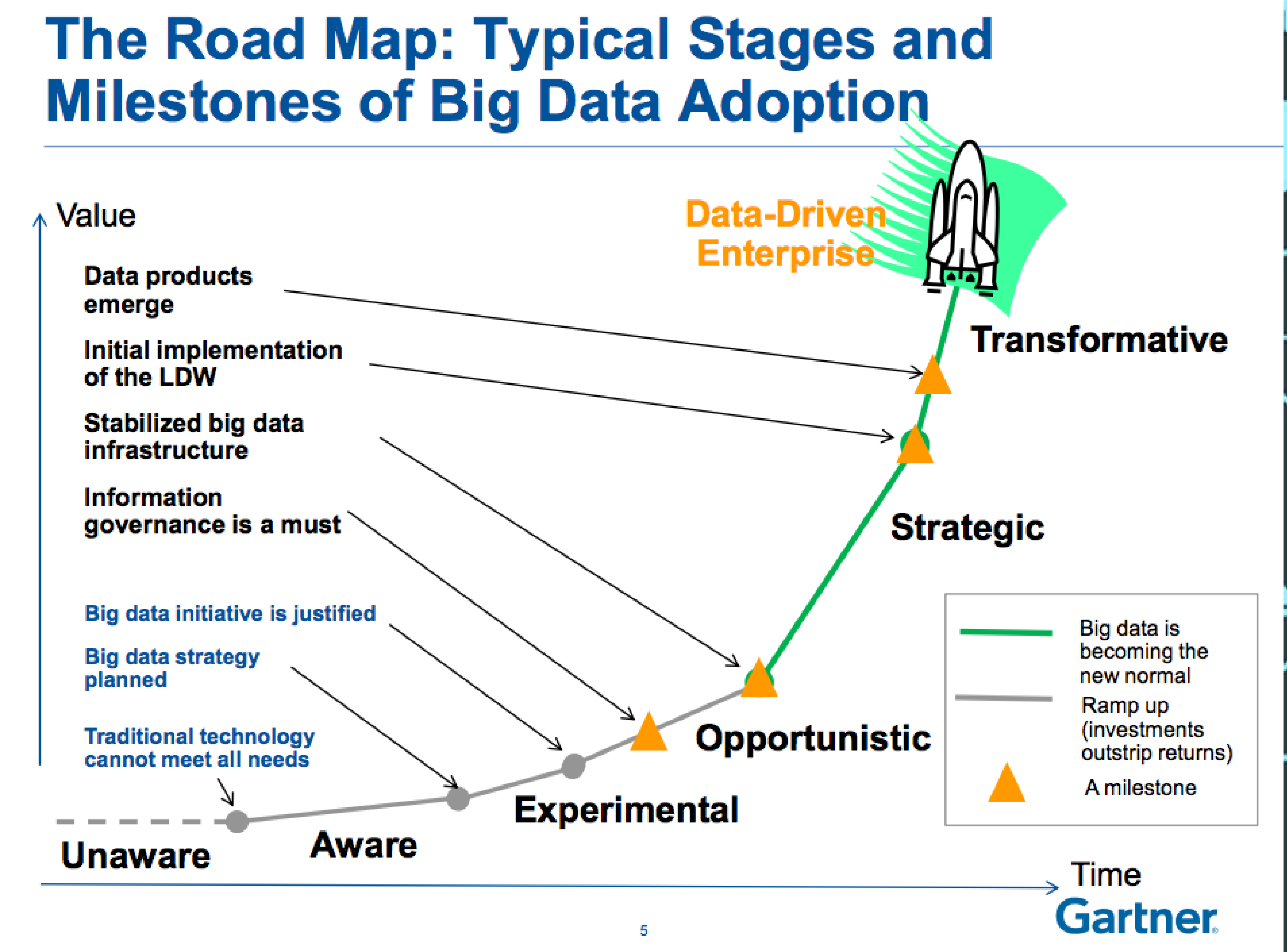
Note: Of the 25 services covered in our survey, 15 are shown here. Usage means that a user completed at least some part of a transaction online. Each country may not allow end-to-end transactions online for a service. Survey question (usage): "Have you used the internet for the following interactions with government at least once in the past two years?" Respondents who answered "Yes" have been included. Survey question (satisfaction): "How satisfied are you with the use of the internet in delivering each kind of government service?" Response options range from 1 to 7, where 1 = "Extremely dissatisfied" and 7 = "Extremely satisfied." Respondents who selected 6 or 7 have been included as satisfied.

Большие данные в гос. управлении

- Открытые данные
- Оборона
- Защита прав потребителей
- Общественная безопасность
- Транспорт и городская инфраструктура
- Общественное здравоохранение
- Энергетика, окружающая среда и ЖКХ
- Образование



Внедрение Больших Данных



The Road Map: Typical Stages and Milestones of Big Data Adoption

Value

Data products emerge

Initial implementation of the LDW

Stabilized big data infrastructure

Information governance is a must

Big data initiative is justified

Big data strategy planned

Traditional technology cannot meet all needs

Data-Driven Enterprise

Transformative

Strategic

Opportunistic

Experimental

Aware

Unaware

Time

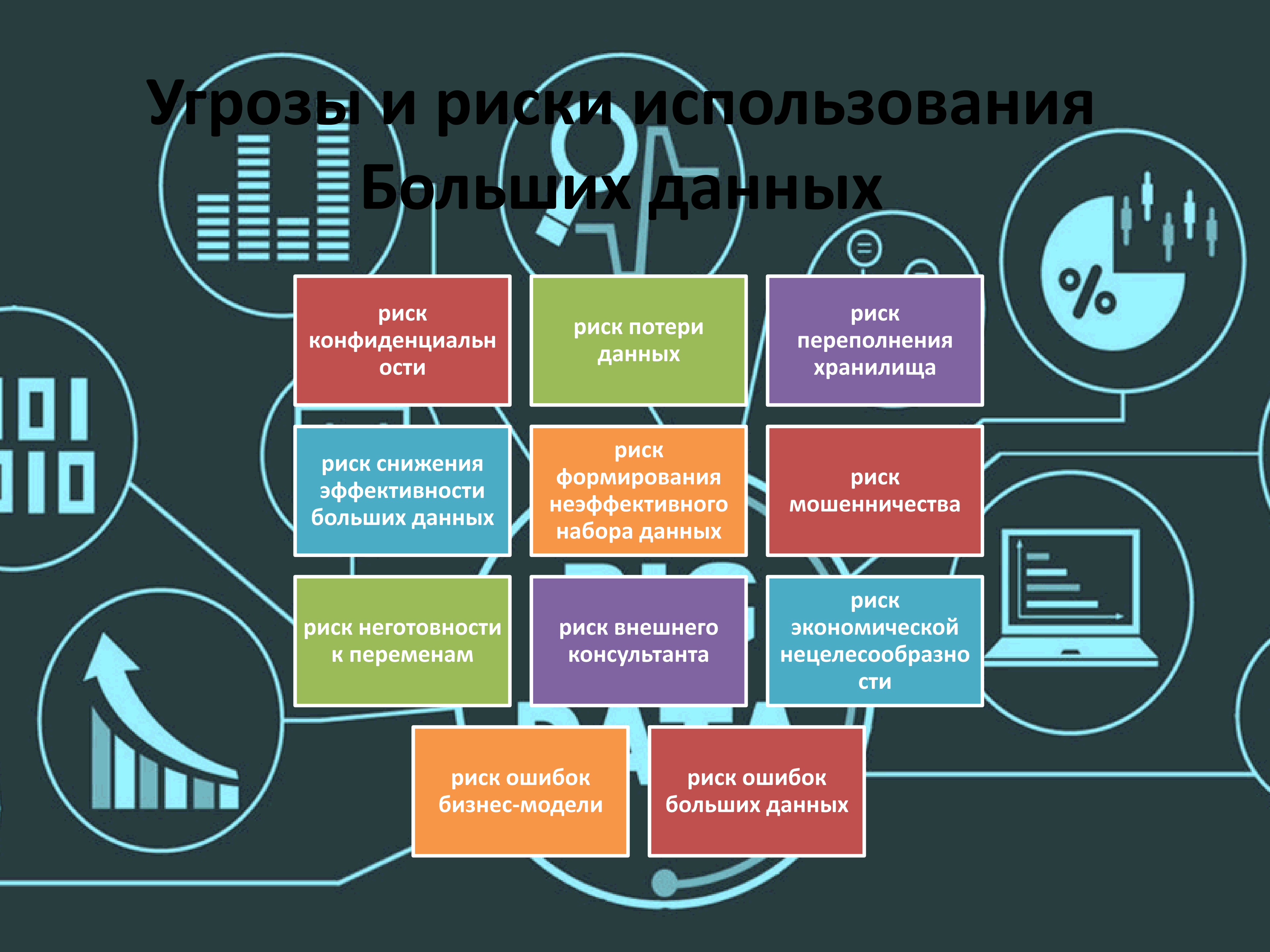
🟩 Big data is becoming the new normal

▭ Ramp up (investments outstrip returns)

▲ A milestone

Gartner

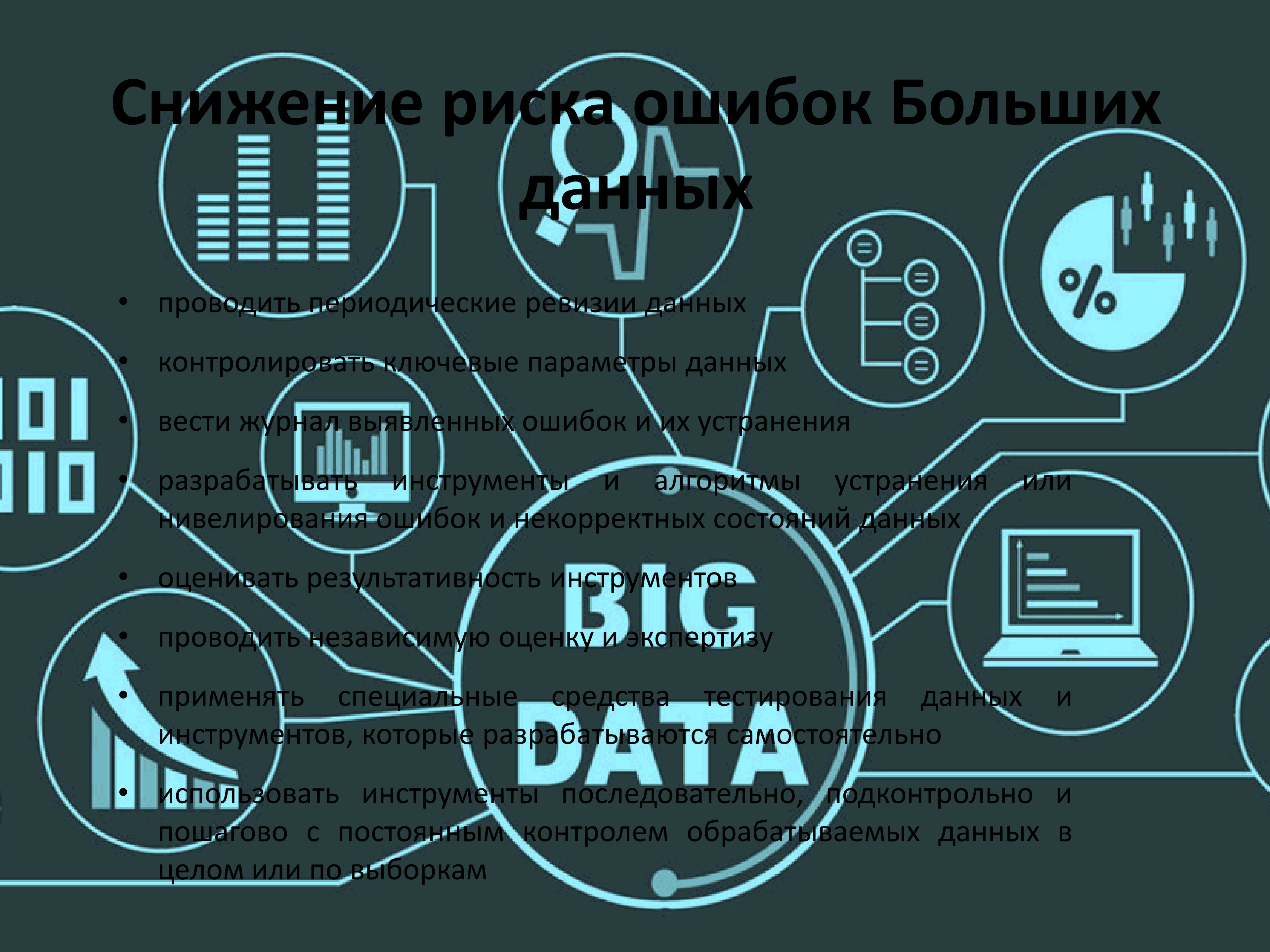
Угрозы и риски использования Больших данных



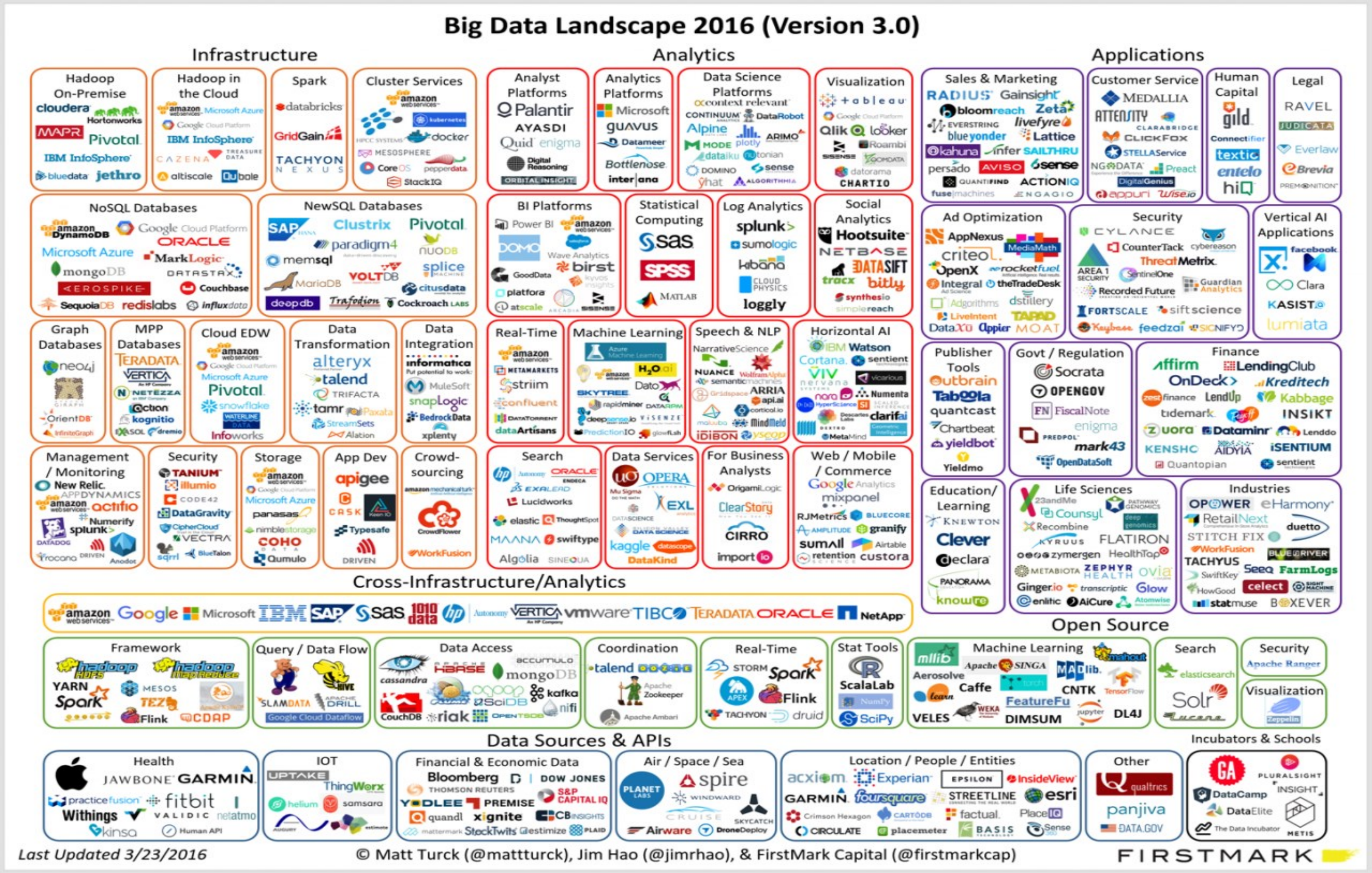
- риск конфиденциальности
- риск потери данных
- риск переполнения хранилища
- риск снижения эффективности больших данных
- риск формирования неэффективного набора данных
- риск мошенничества
- риск неготовности к переменам
- риск внешнего консультанта
- риск экономической нецелесообразности
- риск ошибок бизнес-модели
- риск ошибок больших данных

Снижение риска ошибок Больших данных

- проводить периодические ревизии данных
- контролировать ключевые параметры данных
- вести журнал выявленных ошибок и их устранения
- разрабатывать инструменты и алгоритмы устранения или нивелирования ошибок и некорректных состояний данных
- оценивать результативность инструментов
- проводить независимую оценку и экспертизу
- применять специальные средства тестирования данных и инструментов, которые разрабатываются самостоятельно
- использовать инструменты последовательно, подконтрольно и пошагово с постоянным контролем обрабатываемых данных в целом или по выборкам



Технологии



Big Data Landscape 2016 (Version 3.0)

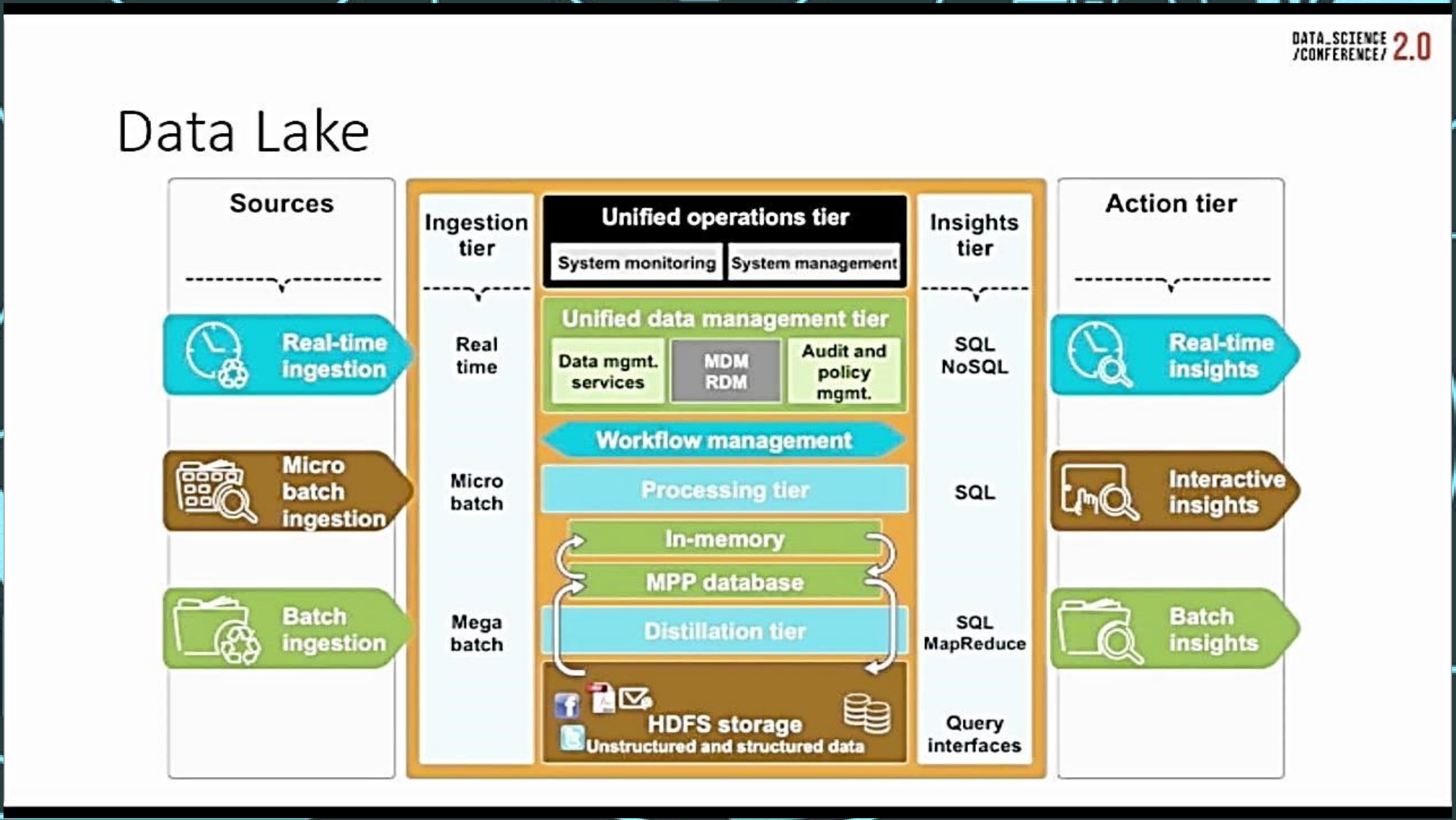
Infrastructure
Analytics
Applications
Hadoop On-Premise: cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, bluedata, jethro
Hadoop in the Cloud: amazon web services, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, Treasure Data, altiscale, Qubole
Spark: databricks, GridGain, TACHYON NEXUS
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, Bottlenose, interana
Data Science Platforms: context relevant, Continuum Analytics, DataRobot, Alpine, ARIMO, MODE, plotly, dataiku, nutonian, DOMINO, Sense, yhat, Algorithmia
Visualization: tableau, Google Cloud Platform, Qlik, looker, SISENSE, Roambi, datorama, CHARTIO
Sales & Marketing: RADIUS, Gainsight, bloomreach, Zeta, EverString, livefyre, blue yonder, Lattice, kahuna, infer, SAILTHRU, persado, AVISO, 6sense, QuantiFind, ACTIONIQ, fuse machines, ENGAGIO
Customer Service: MEDALLIA, ATTENSITY, Clarabridge, ClickFox, STELLAService, NGDATA, Preact, DigitalGenius, appuri, Wise.io
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, Sequoia DB, redislabs, influxdata
NewSQL Databases: SAP HANA, Clustrix, Pivotal, paradigm4, NuoDB, memsql, VoltDB, splice, MariaDB, citusdata, deepdb, Trafodion, Cockroach Labs
BI Platforms: Power BI, amazon web services, DOMO, Wave Analytics, birst, GoodData, platfora, atscale, Arcadia, Sisense
Statistical Computing: sas, SPSS, MATLAB
Log Analytics: splunk, sumologic, kibana, Cloud Physics, loggly
Social Analytics: Hootsuite, NETBASE, DATASIFT, tracx, bitly, synthesio, simplereach
Ad Optimization: AppNexus, MediaMath, criteo, OpenX, rocketfuel, Integral Ad Science, theTradeDesk, Adgorithms, dstillery, Liveintent, TAPAD, DataXu, Appier, MOAT
Security: CYLANCE, CounterTack, cybereason, ThreatMetrix, AREA 1 Security, SentinelOne, Guardian Analytics, Recorded Future, FORTSCALE, sift science, Kaybase, feedzai, SIGNIFYD
Vertical AI Applications: X.ai, facebook, Clara, KASISTO, lumiata
Graph Databases: neo4j, Graph, OrientDB, InfiniteGraph
MPP Databases: TERADATA, VERTICA, NETEZZA, Actian, kognitio, EXASOL, dremio
Cloud EDW: amazon web services, Google Cloud Platform, Microsoft Azure, Pivotal, snowflake, Waterline Data, Infoworks
Data Transformation: alteryx, talend, TRIFACTA, tamr, Paxata, StreamSets, Alation
Data Integration: informatica, MuleSoft, snapLogic, Bedrock Data, xplenty
Real-Time: amazon web services, METAMARKETS, striim, 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, NUANCE, Wolfram Alpha, semantic machines, ARRIA, Gridspace, api.ai, maluuba, MindMeld, idibon, yseop
Horizontal AI: IBM Watson, Cortana, sentient, VIV, nervana, vicarious, nara, Numenta, HyperScience, Descartes Labs, clarifai, MetaMind
Publisher Tools: outbrain, Taboola, quantcast, Chartbeat, yieldbot, Yieldmo
Govt / Regulation: Socrata, OPENGOV, FiscalNote, enigma, PREDPOL, mark43, OpenDataSoft
Finance: affirm, LendingClub, OnDeck, Kreditech, zestfinance, LendUp, Kabbage, tidemark, INSIKT, Zuora, Dataminr, Lenddo, KENSHO, AIDYIA, iSENTIUM, Quantopian, sentient
Management / Monitoring: New Relic, APPDYNAMICS, amazon web services, actifio, Numerify, splunk, DATADOG, Ircona, DRIVEN, Anodot
Security: TANIUM, illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrrl, BlueTalon
Storage: amazon web services, Google Cloud Platform, Microsoft Azure, panasas, nimblestorage, COHO DATA, Qumulo
App Dev: apigee, CASK, Keen IO, Typesafe, DRIVEN
Crowdsourcing: amazon mechanical turk, CrowdFlower, WorkFusion
Search: Autonomy, ORACLE ENDECA, EXALEAD, Lucidworks, elastic, ThoughtSpot, MAANA, swiftype, Algolia, SINEQUA
Data Services: Mu Sigma, OPERA, EXL, DATASCIENCE, Silicon Valley Data Science, kaggle, datascope, DataKind
For Business Analysts: OrigamiLogic, ClearStory, CIRRO, import.io
Web / Mobile / Commerce: Google Analytics, mixpanel, RJMetrics, BLUECORE, AMPLITUDE, granify, sumAll, Airtable, retention science, custora
Education / Learning: KNEWTON, Clever, declara, PANORAMA, knowre
Life Sciences: 23andMe, Counsyl, Pathway Genomics, Recombine, deep genomics, FLATIRON, KYRUUS, 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
Cross-Infrastructure/Analytics: amazon web services, Google, Microsoft, IBM, SAP, sas, 1010data, hp, Autonomy, VERTICA, vmware, TIBCO, TERADATA, ORACLE, NetApp
Open Source
Framework: hadoop HDFS, hadoop MapReduce, YARN, MESOS, Spark, TEZ, Flink, CDAP
Query / Data Flow: pig, HIVE, SLAMDATA, APACHE DRILL, Google Cloud Dataflow
Data Access: cassandra, HBASE, accumulo, mongoDB, kafka, ALLUXIO, SciDB, nifi, CouchDB, riak, OpenTSDB
Coordination: talend, OOZIE, Apache Zookeeper, Apache Ambari
Real-Time: STORM, Spark, APEX, Flink, TACHYON, druid
Stat Tools: R, ScalaLab, NumPy, SciPy
Machine Learning: mllib, mahout, Apache SINGA, MADlib, Aerosolve, torch, Caffe, CNTK, TensorFlow, learn, WEKA, FeatureFu, VELES, DIMSUM, jupyter, DL4J
Search: elasticsearch, Solr, Lucene
Security: Apache Ranger
Visualization: Zeppelin
Data Sources & APIs
Health: JAWBONE, GARMIN, practicefusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API
IOT: UPTAKE, ThingWorx, helium, samsara, AUGURY, estimote
Financial & Economic Data: Bloomberg, DOW JONES, THOMSON REUTERS, S&P CAPITAL IQ, YODLEE, PREMISE, quandl, xignite, CB Insights, mattermark, StockTwits, estimize, PLAID
Air / Space / Sea: PLANET LABS, spire, WINDWARD, CRUISE, SKYCATCH, Airware, DroneDeploy
Location / People / Entities: acxiom, Experian, EPSILON, InsideView, GARMIN, foursquare, STREETLINE, esri, Crimson Hexagon, CARTODB, factual, PlaceIQ, CIRCULATE, placemeter, BASIS, Sense
Other: qualtrics, panjiva, DATA.GOV
Incubators & Schools: GA, PLURALSIGHT, DataCamp, INSIGHT, DataElite, The Data Incubator, METIS

Last Updated 3/23/2016

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK

Data Lake



DATA_SCIENCE /CONFERENCE/ 2.0

Data Lake

Sources

Real-time ingestion

Micro batch ingestion

Batch ingestion

Ingestion tier

Real time

Micro batch

Mega batch

Unified operations tier

System monitoring

System management

Unified data management tier

Data mgmt. services

MDM RDM

Audit and policy mgmt.

Workflow management

Processing tier

In-memory

MPP database

Distillation tier

HDFS storage

Unstructured and structured data

Insights tier

SQL NoSQL

SQL

SQL MapReduce

Query interfaces

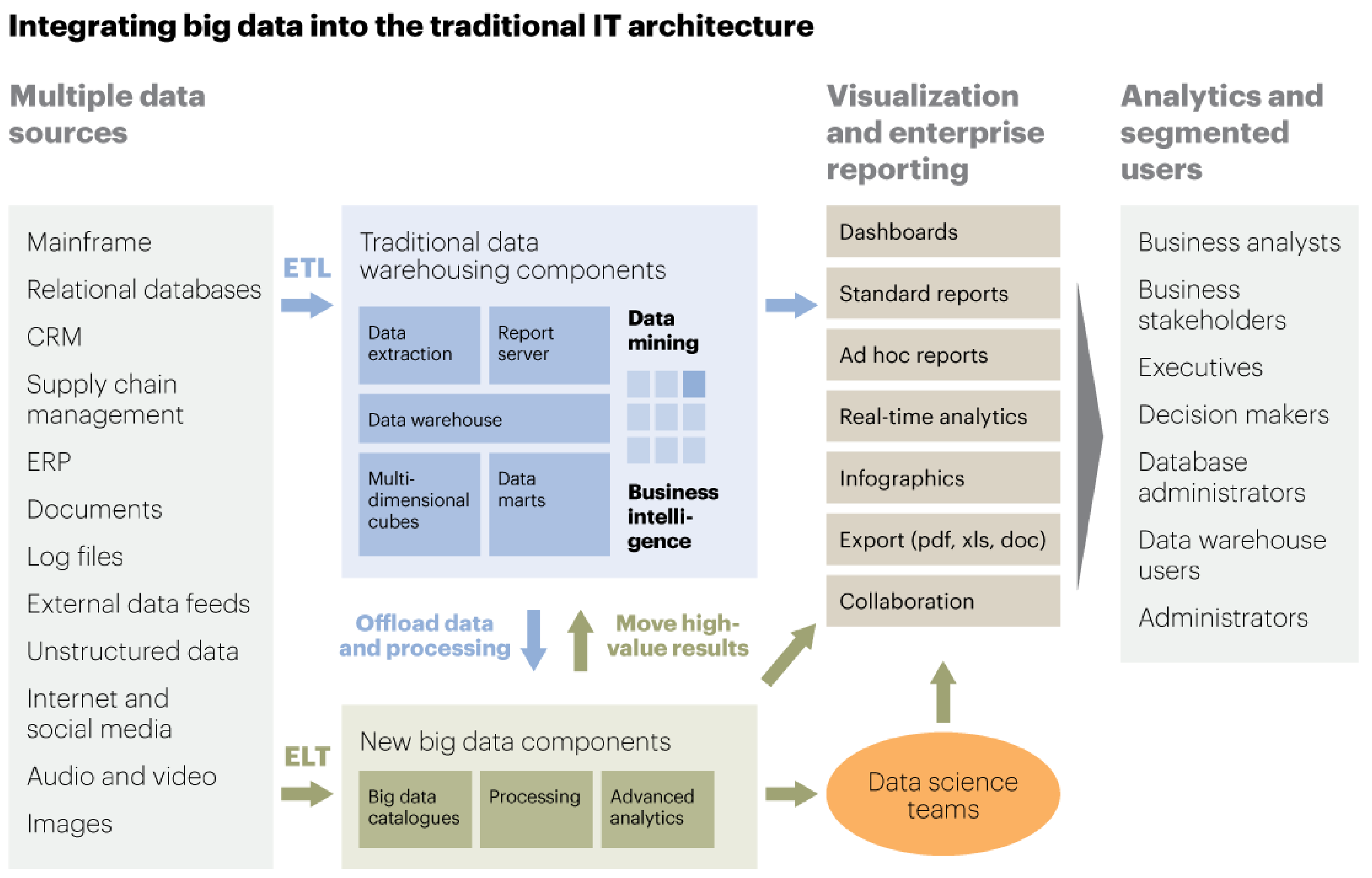
Action tier

Real-time insights

Interactive insights

Batch insights

Процесс работы с данными



Integrating big data into the traditional IT architecture

Multiple data sources

- Mainframe
- Relational databases
- CRM
- Supply chain management
- ERP
- Documents
- Log files
- External data feeds
- Unstructured data
- Internet and social media
- Audio and video
- Images

ETL

Traditional data warehousing components

- Data extraction
- Report server
- Data warehouse
- Multi-dimensional cubes
- Data marts
- **Data mining**
- **Business intelligence**

Offload data and processing

Move high-value results

ELT

New big data components

- Big data catalogues
- Processing
- Advanced analytics

Visualization and enterprise reporting

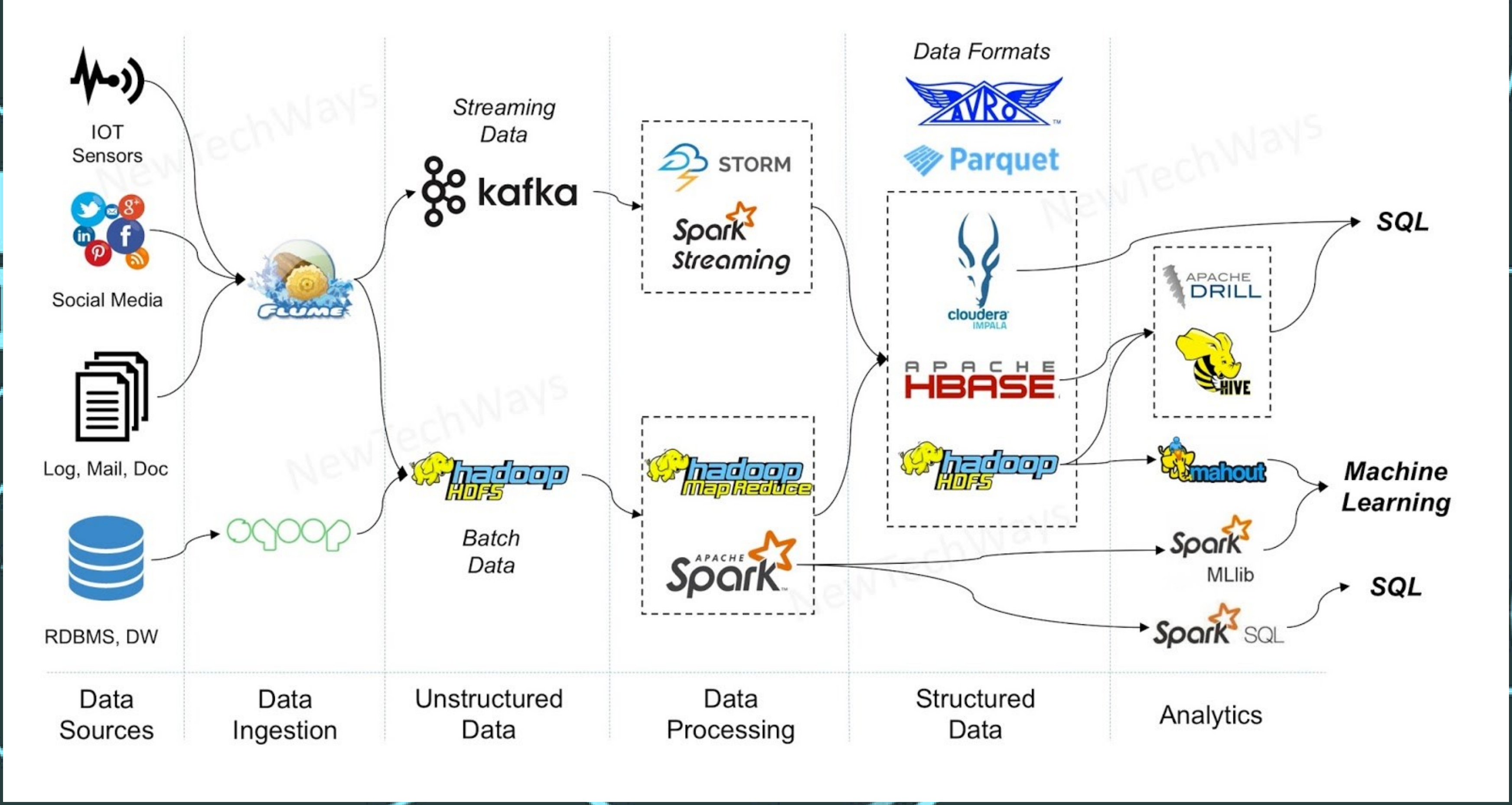
- Dashboards
- Standard reports
- Ad hoc reports
- Real-time analytics
- Infographics
- Export (pdf, xls, doc)
- Collaboration

Data science teams

Analytics and segmented users

- Business analysts
- Business stakeholders
- Executives
- Decision makers
- Database administrators
- Data warehouse users
- Administrators

Процесс Apache



IOT Sensors

Social Media

Log, Mail, Doc

RDBMS, DW

FLUME

sqoop

Streaming Data

kafka

hadoop HDFS

Batch Data

STORM

Spark Streaming

hadoop Map Reduce

APACHE Spark

Data Formats

AVRO

Parquet

cloudera IMPALA

APACHE HBASE

hadoop HDFS

APACHE DRILL

HIVE

mahout

Spark MLlib

Spark SQL

SQL

Machine Learning

SQL

Data Sources

Data Ingestion

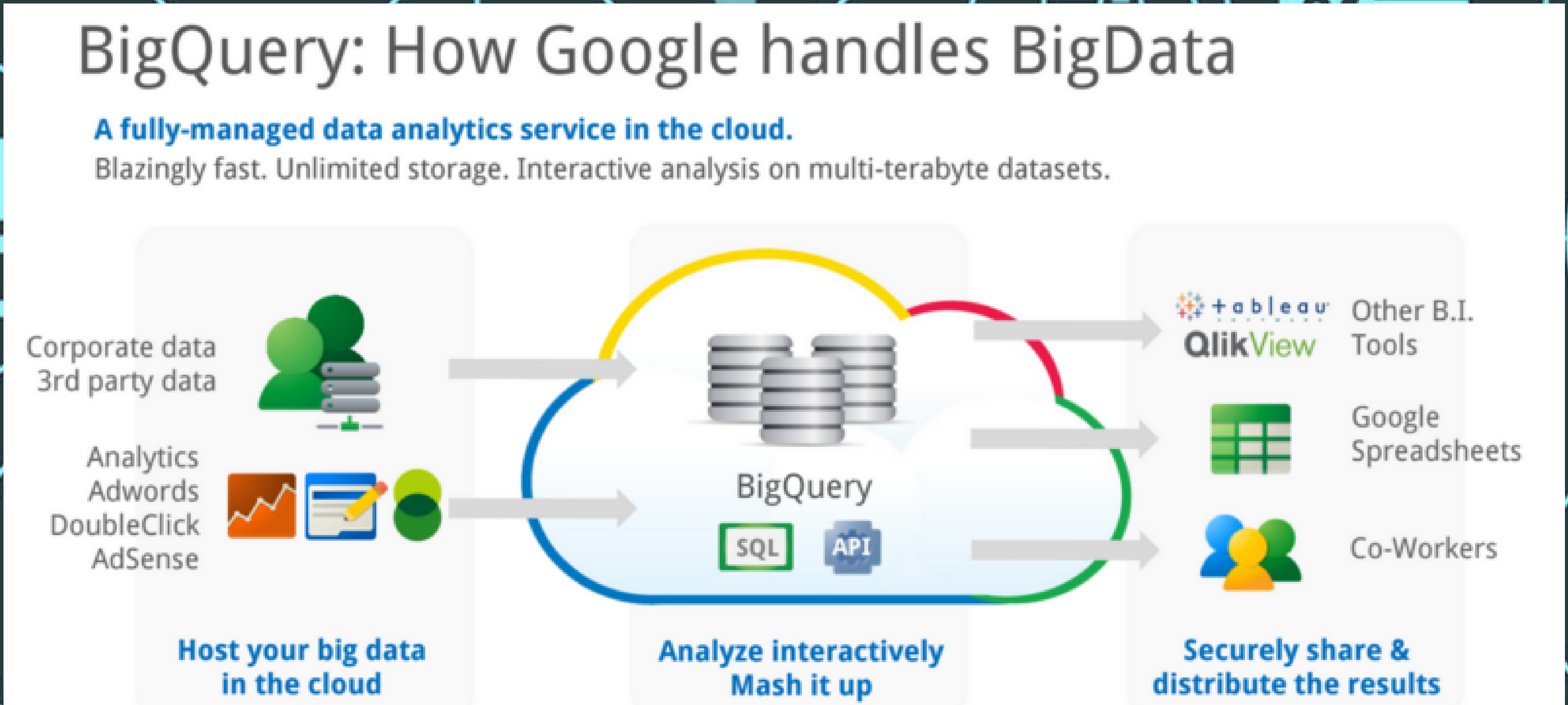
Unstructured Data

Data Processing

Structured Data

Analytics

Процесс Google



BigQuery: How Google handles BigData

A fully-managed data analytics service in the cloud.

Blazingly fast. Unlimited storage. Interactive analysis on multi-terabyte datasets.

Corporate data
3rd party data

Analytics
Adwords
DoubleClick
AdSense

BigQuery

SQL

API

tableau
QlikView

Other B.I.
Tools

Google
Spreadsheets

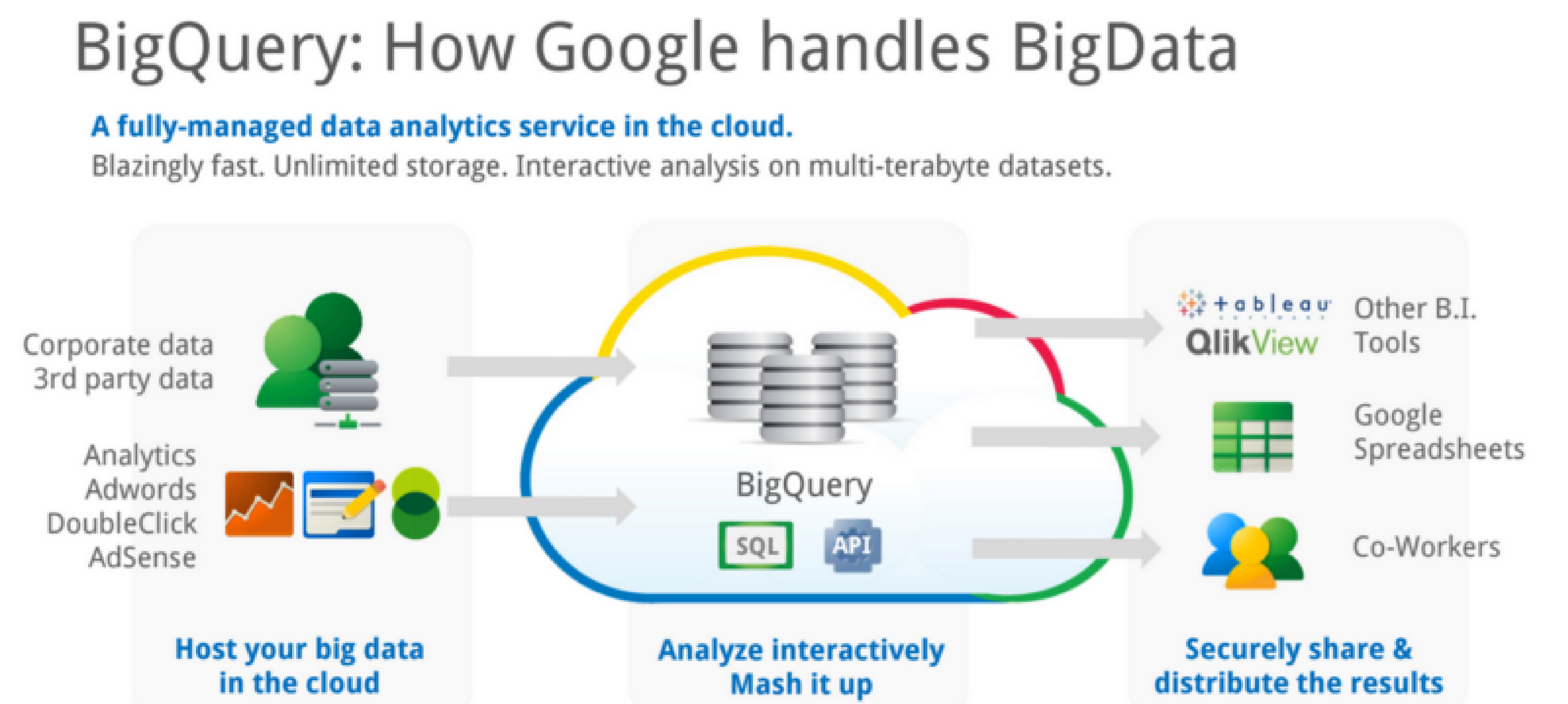
Co-Workers

Host your big data in the cloud

Analyze interactively Mash it up

Securely share & distribute the results

Процесс Google



BigQuery: How Google handles BigData

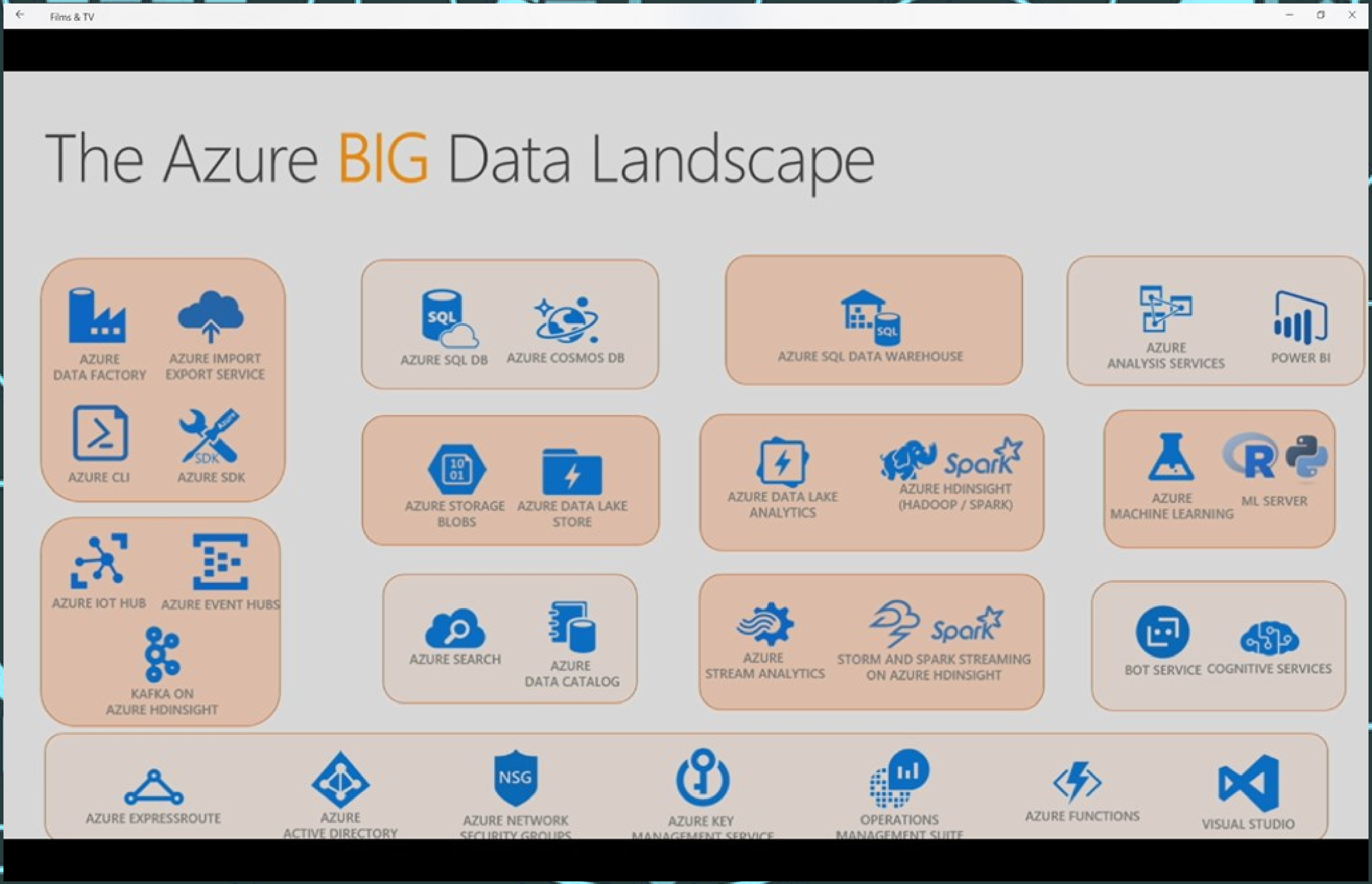
A fully-managed data analytics service in the cloud.
Blazingly fast. Unlimited storage. Interactive analysis on multi-terabyte datasets.

Host your big data in the cloud

Analyze interactively Mash it up

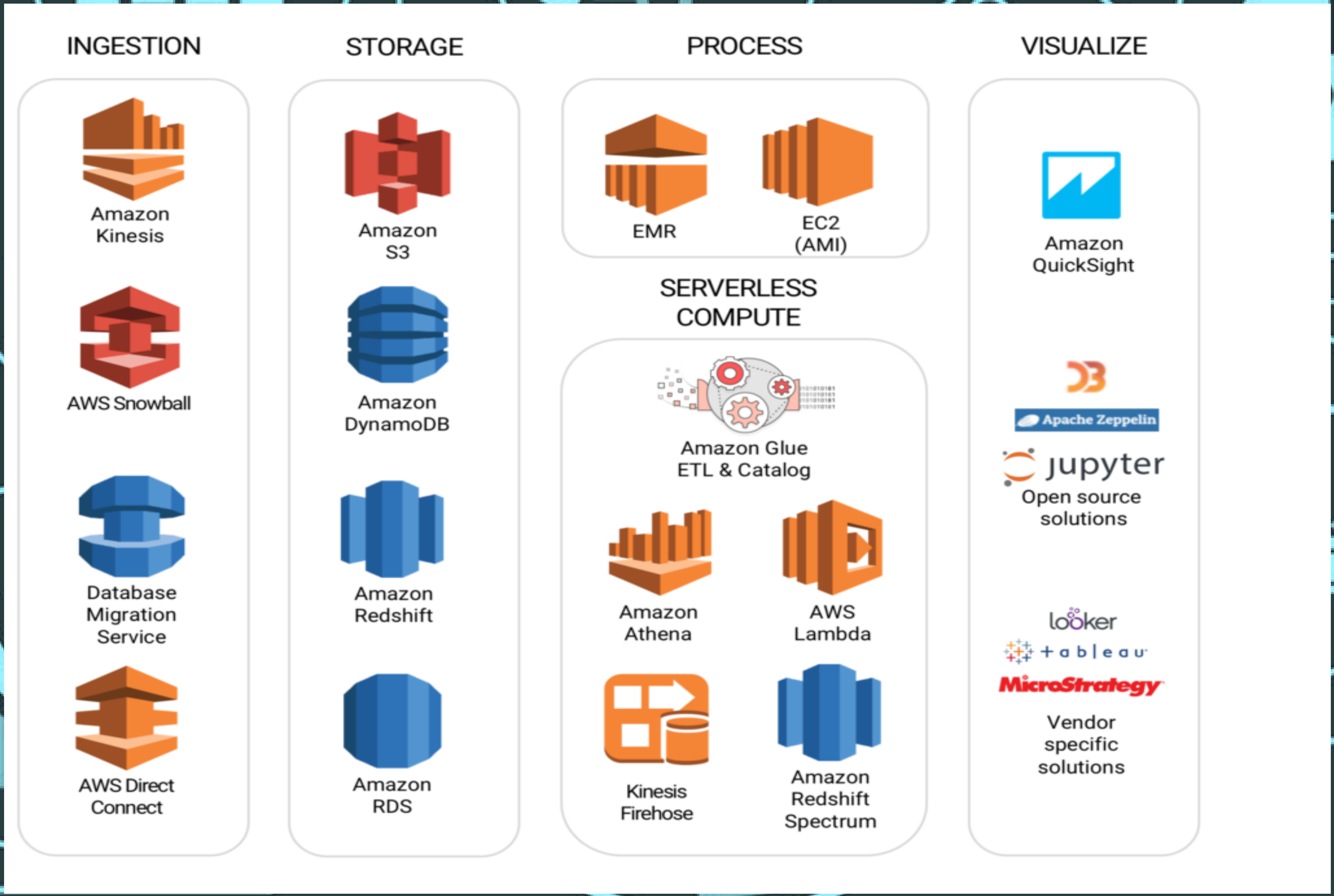
Securely share & distribute the results

Процесс Microsoft



The Azure BIG Data Landscape

Процесс Amazon



INGESTION

Amazon Kinesis

AWS Snowball

Database Migration Service

AWS Direct Connect

STORAGE

Amazon S3

Amazon DynamoDB

Amazon Redshift

Amazon RDS

PROCESS

EMR

EC2 (AMI)

SERVERLESS COMPUTE

Amazon Glue ETL & Catalog

Amazon Athena

AWS Lambda

Kinesis Firehose

Amazon Redshift Spectrum

VISUALIZE

Amazon QuickSight

D3

Apache Zeppelin

jupyter

Open source solutions

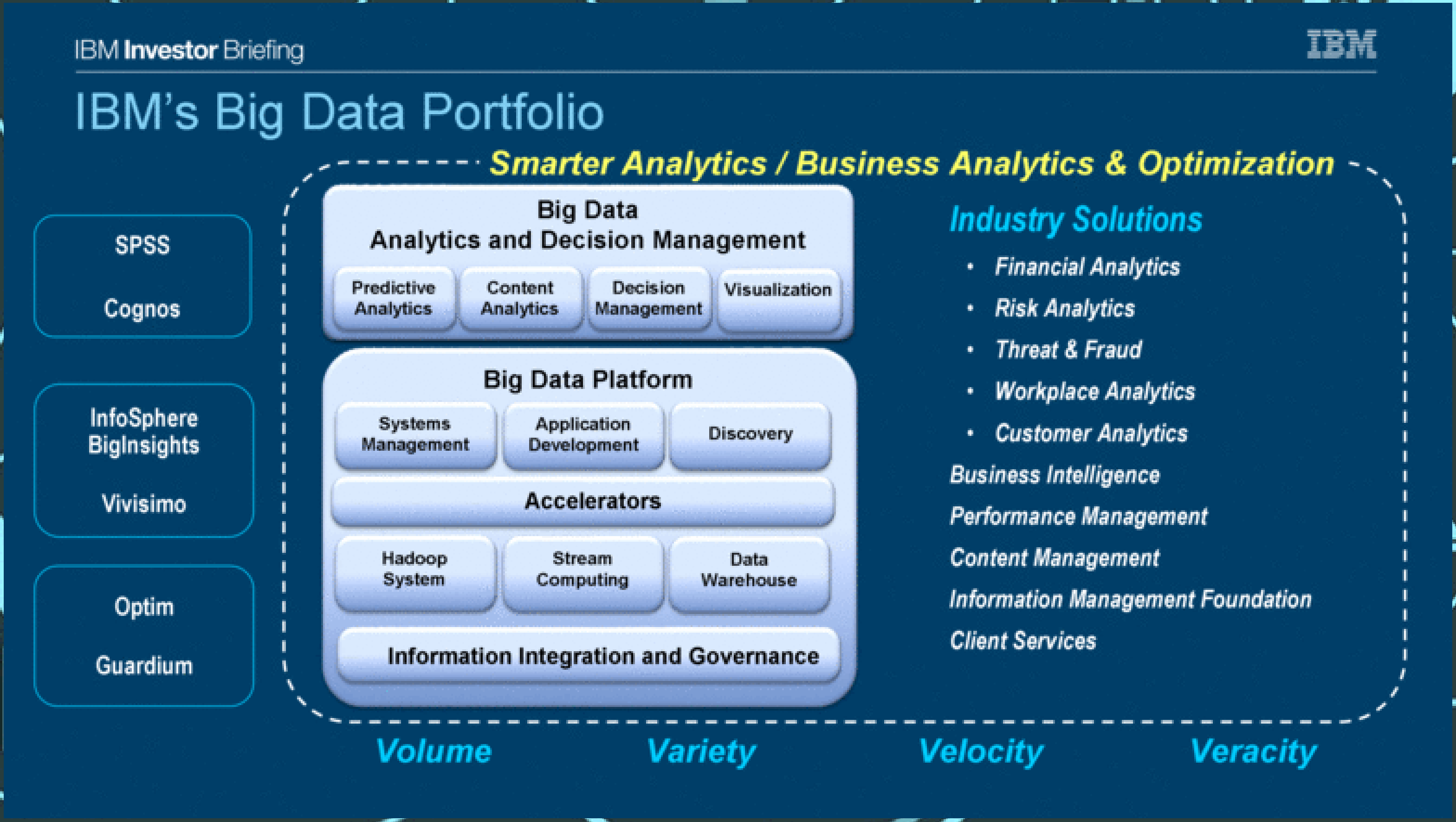
looker

tableau

MicroStrategy

Vendor specific solutions

Процесс IBM



IBM **Investor** Briefing

IBM

IBM's Big Data Portfolio

Smarter Analytics / Business Analytics & Optimization

SPSS

Cognos

InfoSphere BigInsights

Vivisimo

Optim

Guardium

Big Data Analytics and Decision Management

- Predictive Analytics
- Content Analytics
- Decision Management
- Visualization

Big Data Platform

- Systems Management
- Application Development
- Discovery

Accelerators

- Hadoop System
- Stream Computing
- Data Warehouse

Information Integration and Governance

Industry Solutions

- *Financial Analytics*
- *Risk Analytics*
- *Threat & Fraud*
- *Workplace Analytics*
- *Customer Analytics*

Business Intelligence

Performance Management

Content Management

Information Management Foundation

Client Services

Volume *Variety* *Velocity* *Veracity*