Data And Analytics Glossary

ACID Test

A Test Applied To Data For Atomicity, Consistency, Isolation, And Durability.

Ad Hoc Reporting

Reports Generated For A One-Time Need.

Ad Targeting

The Attempt To Reach A Specific Audience With A Specific Message, Typically By Either Contacting Them Directly Or Placing Contextual Ads On The Web.

Algorithm

A Mathematical Formula Placed In Software That Performs An Analysis On A Set Of Data.

Analytics

Using Software-Based Algorithms And Statistics To Derive Meaning From Data.

Analytics Platform

Software Or Software And Hardware That Provides The Tools And Computational Power Needed To Build And Perform Many Different Analytical Queries.

Anonymization

The Severing Of Links Between People In A Database And Their Records To Prevent The Discovery Of The Source Of The Records.

Application

Software That Is Designed To Perform A Specific Task Or Suite Of Tasks.

Automatic Identification And Capture (AIDC)

Any Method Of Automatically Identifying And Collecting Data On Items, And Then Storing The Data In A Computer System. For Example, A Scanner Might Collect Data About A Product Being Shipped Via An RFID Chip.

Behavioral Analytics

Using Data About People’s Behavior To Understand Intent And Predict Future Actions.

Big Data

This Term Has Been Defined In Many Ways, But Along Similar Lines. Doug Laney, Then An Analyst At The META Group, First Defined Big Data In A 2001 Report Called “3 -D Data Management: Controlling Data Volume, Velocity And Variety.” Volume Refers To The Sheer Size Of The Datasets. The Mckinsey Report, “Big Data: The Next Frontier For Innovation, Competition, And Productivity,” Expands On The Volume Aspect By Saying

That, “’Big Data’ Refers To Datasets Whose Size Is Beyond The Ability Of Typical Database Software Tools To Capture, Store, Manage, And Analyze.”

Velocity Refers To The Speed At Which The Data Is Acquired And Used. Not Only Are Companies And Organizations Collecting More And More Data At A Faster Rate, They Want To Derive Meaning From That Data As Soon As Possible, Often In Real Time.

Variety Refers To The Different Types Of Data That Are Available To Collect And Analyze In Addition To The Structured Data Found In A Typical Database. Barry Devlin Of 9sight Consulting Identifies Four Categories Of Information That Constitute Big Data:

1. Machine-Generated Data. This Includes RFID Data, Geolocation Data From Mobile Devices, And Data From Monitoring Devices Such As Utility Meters.

2. Computer Log Data, Such As Clickstreams From Websites.

3. Textual Social Media Information From Sources Such As Twitter And Facebook.

4. Multimedia Social And Other Information From Flickr, Youtube, And Other Similar Sites.

IDC Analyst Benjamin Woo Has Added A Fourth V To The Definition: Value. He Says That Because Big Data Is About Supporting Decisions, You Need The Ability To Act On The Data And Derive Value.

Biometrics

The Use Of Technology To Identify People By One Or More Of Their Physical Traits.

Brand Monitoring

The Act Of Monitoring Your Brand’s Reputation Online, Typically By Using Software To Automate The Process.

Business Intelligence (BI)

The General Term Used For The Identification, Extraction, And Analysis Of Data.

Call Detail Record (CDR) Analysis

Cdrs Contain Data That A Telecommunications Company Collects About Phone Calls, Such As Time And Length Of Call. This Data Can Be Used In Any Number Of Analytical Applications.

Cassandra

A Popular Choice Of Columnar Database For Use In Big Data Applications. It Is An Open Source Database Managed By The Apache Software Foundation.

Cell Phone Data

Cell Phones Generate A Tremendous Amount Of Data, And Much Of It Is Available For Use

With Analytical Applications.

Clickstream Analytics

The Analysis Of Users' Web Activity Through The Items They Click On A Page.

Clojure

Clojure Is A Dynamic Programming Language Based On LISP That Uses The Java Virtual Machine (JVM). It Is Well Suited For Parallel Data Processing.

Cloud

A Broad Term That Refers To Any Internet-Based Application Or Service That Is Hosted Remotely.

Columnar Database Or Column-Oriented Database

A Database That Stores Data By Column Rather Than By Row. In A Row-Based Database, A Row Might Contain A Name, Address, And Phone Number. In A Column-Oriented Database, All Names Are In One Column, Addresses In Another, And So On. A Key Advantage Of A Columnar Database Is Faster Hard Disk Access.

Competitive Monitoring

Keeping Tabs Of Competitors' Activities On The Web Using Software To Automate The Process.

Complex Event Processing (CEP)

CEP Is The Process Of Monitoring And Analyzing All Events Across An Organization’s Systems And Acting On Them When Necessary In Real Time.

Comprehensive Large Array-Data Stewardship System (CLASS)

A Digital Library Of Historical Environmental Data From Satellites Operated By The U.S. National Oceanic And Atmospheric Association (NOAA).

Computer-Generated Data

Any Data Generated By A Computer Rather Than A Human–A Log File For Example.

Concurrency

The Ability To Execute Multiple Processes At The Same Time.

Confabulation

The Act Of Making An Intuition-Based Decision Appear To Be Data-Based.

Content Management System (CMS)

Software That Facilitates The Management And Publication Of Content On The Web.

Cross-Channel Analytics

Analysis That Can Attribute Sales, Show Average Order Value, Or The Lifetime Value.

Crowdsourcing

The Act Of Submitting A Task Or Problem To The Public For Completion Or Solution.

Customer Relationship Management (CRM)

Software That Helps Businesses Manage Sales And Customer Service Processes.

Dashboard

A Graphical Reporting Of Static Or Real-Time Data On A Desktop Or Mobile Device. The Data Represented Is Typically High-Level To Give Managers A Quick Report On Status Or Performance.

Data

A Quantitative Or Qualitative Value. Common Types Of Data Include Sales Figures, Marketing Research Results, Readings From Monitoring Equipment, User Actions On A Website, Market Growth Projections, Demographic Information, And Customer Lists.

Data Access

The Act Or Method Of Viewing Or Retrieving Stored Data.

Data Aggregation

The Act Of Collecting Data From Multiple Sources For The Purpose Of Reporting Or Analysis.

Data Analytics

The Application Of Software To Derive Information Or Meaning From Data. The End Result Might Be A Report, An Indication Of Status, Or An Action Taken Automatically Based On The Information Received.

Data Analyst

A Person Responsible For The Tasks Of Modeling, Preparing, And Cleaning Data For The Purpose Of Deriving Actionable Information From It.

Data Architecture And Design

How Enterprise Data Is Structured. The Actual Structure Or Design Varies Depending On The Eventual End Result Required. Data Architecture Has Three Stages Or Processes: Conceptual Representation Of Business Entities. The Logical Representation Of The Relationships Among Those Entities, And The Physical Construction Of The System To Support The Functionality.

Database

A Digital Collection Of Data And The Structure Around Which The Data Is Organized. The Data Is Typically Entered Into And Accessed Via A Database Management System (DBMS).

Database Administrator (DBA)

A Person, Often Certified, Who Is Responsible For Supporting And Maintaining The Integrity Of The Structure And Content Of A Database.

Database As A Service (Daas)

A Database Hosted In The Cloud And Sold On A Metered Basis. Examples Include Heroku Postgres And Amazon Relational Database Service.

Database Management System (DBMS)

Software That Collects And Provides Access To Data In A Structured Format.

Data Center

A Physical Facility That Houses A Large Number Of Servers And Data Storage Devices. Data Centers Might Belong To A Single Organization Or Sell Their Services To Many Organizations.

Data Cleansing

The Act Of Reviewing And Revising Data To Remove Duplicate Entries, Correct Misspellings, Add Missing Data, And Provide More Consistency.

Data Collection

Any Process That Captures Any Type Of Data.

Data Custodian

A Person Responsible For The Database Structure And The Technical Environment, Including The Storage Of Data.

Data-Directed Decision Making

Using Data To Support Making Crucial Decisions.

Data Exhaust

The Data That A Person Creates As A Byproduct Of A Common Activity–For Example, A Cell Call Log Or Web Search History.

Data Feed

A Means For A Person To Receive A Stream Of Data. Examples Of Data Feed Mechanisms Include RSS Or Twitter.

Data Governance

A Set Of Processes Or Rules That Ensure The Integrity Of The Data And That Data Management Best Practices Are Met.

Data Integration

The Process Of Combining Data From Different Sources And Presenting It In A Single View.

Data Integrity

The Measure Of Trust An Organization Has In The Accuracy, Completeness, Timeliness, And

Validity Of The Data.

Data Management

According To The Data Management Association, Data Management Incorporates The Following Practices Needed To Manage The Full Data Lifecycle In An Enterprise:

- Data Governance
- Data Architecture, Analysis, And Design
- Database Management
- Data Security Management
- Data Quality Management
- Reference And Master Data Management
- Data Warehousing And Business Intelligence Management
- Document, Record, And Content Management
- Metadata Management
- Contact Data Management

Data Management Association (DAMA)

A Non-Profit International Organization For Technical And Business Professionals “Dedicated To Advancing The Concepts And Practices Of Information And Data Management.”

Data Marketplace

A Place Where People Can Buy And Sell Data Online.

Data Mart

The Access Layer Of A Data Warehouse Used To Provide Data To Users.

Data Migration

The Process Of Moving Data Between Different Storage Types Or Formats, Or Between Different Computer Systems.

Data Mining

The Process Of Deriving Patterns Or Knowledge From Large Data Sets.

Data Model, Data Modeling

A Data Model Defines The Structure Of The Data For The Purpose Of Communicating Between Functional And Technical People To Show Data Needed For Business Processes, Or For Communicating A Plan To Develop How Data Is Stored And Accessed Among Application Development Team Members.

Data Point

An Individual Item On A Graph Or A Chart.

Data Profiling

The Process Of Collecting Statistics And Information About Data In An Existing Source.

Data Quality

The Measure Of Data To Determine Its Worthiness For Decision Making, Planning, Or Operations.

Data Replication

The Process Of Sharing Information To Ensure Consistency Between Redundant Sources.

Data Repository

The Location Of Permanently Stored Data.

Data Science

A Recent Term That Has Multiple Definitions, But Generally Accepted As A Discipline That Incorporates Statistics, Data Visualization, Computer Programming, Data Mining, Machine Learning, And Database Engineering To Solve Complex Problems.

Data Scientist

A Practitioner Of Data Science.

Data Security

The Practice Of Protecting Data From Destruction Or Unauthorized Access.

Data Set

A Collection Of Data, Typically In Tabular Form.

Data Source

Any Provider Of Data–For Example, A Database Or A Data Stream.

Data Steward

A Person Responsible For Data Stored In A Data Field.

Data Structure

A Specific Way Of Storing And Organizing Data.

Data Visualization

A Visual Abstraction Of Data Designed For The Purpose Of Deriving Meaning Or Communicating Information More Effectively.

Data Virtualization

The Process Of Abstracting Different Data Sources Through A Single Data Access Layer.

Data Warehouse

A Place To Store Data For The Purpose Of Reporting And Analysis.

De-Identification

The Act Of Removing All Data That Links A Person To A Particular Piece Of Information.

Demographic Data

Data Relating To The Characteristics Of A Human Population.

Deep Thunder

IBM's Weather Prediction Service That Provides Weather Data To Organizations Such As Utilities, Which Use The Data To Optimize Energy Distribution.

Distributed Cache

A Data Cache That Is Spread Across Multiple Systems But Works As One. It Is Used To Improve Performance.

Distributed Object

A Software Module Designed To Work With Other Distributed Objects Stored On Other Computers.

Distributed Processing

The Execution Of A Process Across Multiple Computers Connected By A Computer Network.

Document Management

The Practice Of Tracking And Storing Electronic Documents And Scanned Images Of Paper Documents.

Drill

An Open Source Distributed System For Performing Interactive Analysis On Large-Scale Datasets. It Is Similar To Google's Dremel, And Is Managed By Apache.

Elasticsearch

An Open Source Search Engine Built On Apache Lucene.

Electronic Health Records (EHR)

A Digitized Health Record Meant To Be Usable Across Different Health Care Settings.

Enterprise Resource Planning (ERP)

A Software System That Allows An Organization To Coordinate And Manage All Its Resources, Information, And Business Functions.

Event Analytics

Shows The Series Of Steps That Led To An Action.

Exabyte

One Million Terabytes, Or 1 Billion Gigabytes Of Information.

External Data

Data That Exists Outside Of A System.

Extract, Transform, And Load (ETL)

A Process Used In Data Warehousing To Prepare Data For Use In Reporting Or Analytics.

Failover

The Automatic Switching To Another Computer Or Node Should One Fail.

Federal Information Security Management Act (FISMA)

A US Federal Law That Requires All Federal Agencies To Meet Certain Standards Of Information Security Across Its Systems.

Grid Computing

The Performing Of Computing Functions Using Resources From Multiple Distributed Systems. Grid Computing Typically Involves Large Files And Are Most Often Used For Multiple Applications. The Systems That Comprise A Grid Computing Network Do Not Have To Be Similar In Design Or In The Same Geographic Location.

Hadoop

An Open Source Software Library Project Administered By The Apache Software Foundation. Apache Defines Hadoop As “A Framework That Allows For The Distributed Processing Of Large Data Sets Across Clusters Of Computers Using A Simple Programming Model.”

HANA

A Software/Hardware In-Memory Computing Platform From SAP Designed For High-Volume Transactions And Real-Time Analytics.

Hbase

A Distributed Columnar Nosql Database.

High-Performance Computing (HPC)

HPC Systems, Also Called Supercomputers, Are Often Custom Built From State-Of-The-Art Technology To Maximize Compute Performance, Storage Capacity And Throughput, And Data Transfer Speeds.

Hive

A SQL-Like Query And Data Warehouse Engine.

In-Database Analytics

The Integration Of Data Analytics Into The Data Warehouse.

Information Management

The Practice Of Collecting, Managing, And Dsitributing Information Of All Types–Digital, Paper-Based, Structured, Unstructured.

In-Memory Database

Any Database System That Relies On Memory For Data Storage.

In-Memory Data Grid (IMDG)

The Storage Of Data In Memory Across Multiple Servers For The Purpose Of Greater

Scalability And Faster Access Or Analytics.

Kafka

Linkedin’s Open-Source Message System Used To Monitor Activity Events On The Web.

Latency

Any Delay In A Response Or Delivery Of Data From One Point To Another.

Legacy System

Any Computer System, Application, Or Technology That Is Obsolete, But Continues To Be Used Because It Performs A Needed Function Adequately.

Linked Data

As Described By World Wide Web Inventor Time Berners-Lee, “Cherry-Picking Common Attributes Or Languages To Identify Connections Or Relationships Between Disparate Sources Of Data.”

Load Balancing

The Process Of Distributing Workload Across A Computer Network Or Computer Cluster To Optimize Performance.

Location Analytics

Location Analytics Brings Mapping And Map-Driven Analytics To Enterprise Business Systems And Data Warehouses. It Allows You To Associate Geospatial Information With Datasets.

Location Data

Data That Describes A Geographic Location.

Log File

A File That A Computer, Network, Or Application Creates Automatically To Record Events That Occur During Operation–For Example, The Time A File Is Accessed.

Long Data

A Term Coined By Mathematician And Network Scientist Samuel Arbesman That Refers To “Datasets That Have Massive Historical Sweep.”

Machine-Generated Data

Any Data That Is Automatically Created From A Computer Process, Application, Or Other Non-Human Source.

Machine Learning

The Use Of Algorithms To Allow A Computer To Analyze Data For The Purpose Of “Learning” What Action To Take When A Specific Pattern Or Event Occurs.

Map/Reduce

A General Term That Refers To The Process Of Breaking Up A Problem Into Pieces That Are Then Distributed Across Multiple Computers On The Same Network Or Cluster, Or

Across A Grid Of Disparate And Possibly Geographically Separated Systems (Map), And Then Collecting All The Results And Combines Them Into A Report (Reduce). Google's Branded Framework To Perform This Function Is Called Mapreduce.

Mashup

The Process Of Combining Different Datasets Within A Single Application To Enhance Output–For Example, Combining Demographic Data With Real Estate Listings.

Massively Parallel Processing (MPP)

The Act Of Processing Of A Program By Breaking It Up Into Separate Pieces, Each Of Which Is Executed On Its Own Processor, Operating System, And Memory.

Master Data Management (MDM)

Master Data Is Any Non-Transactional Data That Is Critical To The Operation Of A Business–For Example, Customer Or Supplier Data, Product Information, Or Employee Data. MDM Is The Process Of Managing That Data To Ensure Consistency, Quality, And Availability.

Metadata

Any Data Used To Describe Other Data–For Example, A Data File's Size Or Date Of Creation.

Mongodb

An Open-Source Nosql Database Managed By 10gen.

MPP Database

A Database Optimized To Work In A Massively Parallel Processing Environment.

Multi-Threading

The Act Of Breaking Up An Operation Within A Single Computer System Into Multiple Threads For Faster Execution.

Nosql

A Class Of Database Management System That Does Not Use The Relational Model. Nosql Is Designed To Handle Large Data Volumes That Do Not Follow A Fixed Schema. It Is Ideally Suited For Use With Very Large Data Volumes That Do Not Require The Relational Model.

Online Analytical Processing (OLAP)

The Process Of Analyzing Multidimensional Data Using Three Operations: Consolidation (The Aggregation Of Available), Drill-Down (The Ability For Users To See The Underlying

Details), And Slice And Dice (The Ability For Users To Select Subsets And View Them From Different Perspectives).

Online Transactional Processing (OLTP)

The Process Of Providing Users With Access To Large Amounts Of Transactional Data In A Way That They Can Derive Meaning From It.

Opendremel

The Open Source Version Of Google's Big Query Java Code. It Is Being Integrated With Apache Drill.

Open Data Center Alliance (ODCA)

A Consortium Of Global IT Organizations Whose Goal Is To Speed The Migration Of Cloud Computing.

Operational Data Store (ODS)

A Location To Gather And Store Data From Multiple Sources So That More Operations Can Be Performed On It Before Sending To The Data Warehouse For Reporting.

Parallel Data Analysis

Breaking Up An Analytical Problem Into Smaller Components And Running Algorithms On Each Of Those Components At The Same Time. Parallel Data Analysis Can Occur Within The Same System Or Across Multiple Systems.

Parallel Method Invocation (PMI)

Allows Programming Code To Call Multiple Functions In Parallel.

Parallel Processing

The Ability To Execute Multiple Tasks At The Same Time.

Parallel Query

A Query That Is Executed Over Multiple System Threads For Faster Performance.

Pattern Recognition

The Classification Or Labeling Of An Identified Pattern In The Machine Learning Process.

Performance Management

The Process Of Monitoring System Or Business Performance Against Predefined Goals To

Identify Areas That Need Attention.

Petabyte

One Million Gigabytes Or 1,024 Terabytes.

Pig

A Data Flow Language And Execution Framework For Parallel Computation.

Predictive Analytics

Using Statistical Functions On One Or More Datasets To Predict Trends Or Future Events.

Predictive Modeling

The Process Of Developing A Model That Will Most Likely Predict A Trend Or Outcome.

Query Analysis

The Process Of Analyzing A Search Query For The Purpose Of Optimizing It For The Best Possible Result.

R

An Open Source Software Environment Used For Statistical Computing.

Radio-Frequency Identification (RFID)

A Technology That Uses Wireless Communications To Send Information About An Object From One Point To Another.

Real Time

A Descriptor For Events, Data Streams, Or Processes That Have An Action Performed On Them As They Occur.

Recommendation Engine

An Algorithm That Analyzes A Customer's Purchases And Actions On An E-Commerce Site And Then Uses That Data To Recommend Complementary Products.

Records Management

The Process Of Managing An Organization's Records Throughout Their Entire Lifecycle, From Creation To Disposal.

Reference Data

Data That Describes An Object And Its Properties. The Object May Be Physical Or Virtual.

Report

The Presentation Of Information Derived From A Query Against A Dataset, Usually In A Predetermined Format.

Risk Analysis

The Application Of Statistical Methods On One Or More Datasets To Determine The Likely Risk Of A Project, Action, Or Decision.

Root-Cause Analysis

The Process Of Determining The Main Cause Of An Event Or Problem.

Sawzall

Google's Procedural Domain-Specific Programming Language Designed To Process Large Volumes Of Log Records.

Scalability

The Ability Of A System Or Process To Maintain Acceptable Performance Levels As Workload Or Scope Increases.

Schema

The Structure That Defines The Organization Of Data In A Database System.

Search

The Process Of Locating Specific Data Or Content Using A Search Tool.

Search Data

Aggregated Data About Search Terms Used Over Time.

Semantic Web

A Project Of The World Wide Web Consortium (W3C) To Encourage The Use Of A Standard Format To Include Semantic Content On Websites. The Goal Is To Enable Computers And Other Devices To Better Process Data.

Semi-Structured Data

Data That Is Not Structured By A Formal Data Model, But Provides Other Means Of Describing The Data And Hierarchies.

Sentiment Analysis

The Application Of Statistical Functions On Comments People Make On The Web And Through Social Networks To Determine How They Feel About A Product Or Company.

Server

A Physical Or Virtual Computer That Serves Requests For A Software Application And Delivers Those Requests Over A Network.

Smart Grid

The Smart Grid Refers To The Concept Of Adding Intelligence To The World's Electrical Transmission Systems With The Goal Of Optimizing Energy Efficiency. Enabling The Smart Grid Will Rely Heavily On Collecting, Analyzing, And Acting On Large Volumes Of Data.

Smart Meter

An Electrical Meter That Monitor And Report Energy Usage And Are Capable Of Two-Way Communication With The Utility.

Solid-State Drive (SSD)

Also Called A Solid-State Disk, A Device That Uses Memory Ics To Persistently Store Data.

Software As A Service (Saas)

Application Software That Is Used Over The Web By A Thin Client Or Web Browser. Salesforce Is A Well-Known Example Of Saas.

Storage

Any Means Of Storing Data Persistently.

Storm

An Open-Source Distributed Computation System Designed For Processing Multiple Data Streams In Real Time.

Structured Data

Data That Is Organized By A Predetermined Structure.

Structured Query Language (SQL)

A Programming Language Designed Specifically To Manage And Retrieve Data From A Relational Database System.

Terabyte

1,000 Gigabytes.

Text Analytics

The Application Of Statistical, Linguistic, And Machine Learning Techniques On Text-Based Sources To Derive Meaning Or Insight.

Transactional Data

Data That Changes Unpredictably. Examples Include Accounts Payable And Receivable Data, Or Data About Product Shipments.

Transparency

As More Data Becomes Openly Available, The Idea Of Proprietary Data As A Competitive Advantage Is Diminished.

Unstructured Data

Data That Has No Identifiable Structure–For Example, The Text Of Email Messages.

Variable Pricing

The Practice Of Changing Price On The Fly In Response To Supply And Demand. It Requires Real-Time Monitoring Of Consumption And Supply.

Weather Data

Real-Time Weather Data Is Now Widely Available For Organizations To Use In A Variety Of Ways. For Example, A Logistics Company Can Monitor Local Weather Conditions To Optimize The Transport Of Goods. A Utility Company Can Adjust Energy Distribution In Real Time.

Whole Earth Model

An Integrated Data Management System That Allows Geophysicists, Engineers, And Financial Managers In The Oil And Gas Industry Evaluate The Potential Of Oil And Gas Fields.