INCOBRA INFO SESSION SECOND EDITION

'Responsible research and innovation: An opportunity for scientific exchange'

BR-EU project management showcase on Open Data

Carlos Eurico do Canto – CEO Propus Data Science

Porto Alegre, october 27th 2017

Agenda

- Who I am
- About Propus Data Science
- International initiatives
- Open source in Big Data & Data Science Scenario
- Why Open Source?
- Open source benefits
- Open Data

Who I am

Entrepreneur, Electronic Engineer, Computer Scientist, Big Data and Data Science Specialist.

In the last four years dedicated to Big Data, NoSQL and Data Science with several projects developed for companies such as Petrobras, IRB Brasil RE, Zenvia, Grendene.

CEO Propus Data Science.

Propus Data Science

We work to disseminate news concepts and technologies such as NoSQL, Big Data and Data Science.

SERVİCES	TRAININGS	SUPPORT
We provide consultancy in Big Data, NoSQL, Hadoop, Data Science and develop specific projects. We are partners of MongoDB, Neo4J and Hortonworks.	We empower your team. Workshops, in-house training in: MongoDB, Big Data and Hadoop.	Propus Data Science is able to ensure the evolution and operational continuity of it's customer's Big Data structure.

International Initiatives - 2015

Propus Data Science

A brazilian Big Data startup

Propus Science
Carlos Eurico do Canto
CEO

OW2 is an independent, global, open-source software community. The mission of OW2 is to a) promote the development of open-source middleware, generic business applications, cloud computing platforms and b) foster a vibrant community and business ecosystem.

International Initiatives - 2017

MISSÃO TECNOLÓGICA BRASIL - PORTUGAL
"Ponte Luso Brasileira, eterno canal de colaboração, negócios e amizades"

MTECH PORTUGAL 2017

06 e 09 de Novembro
Lisboa, Portugal

1ª Etapa – Brazilian Connection:

Matchmaking, Inteligência, Informações & Networking

Open Source in Big Data & Data Science Scenario
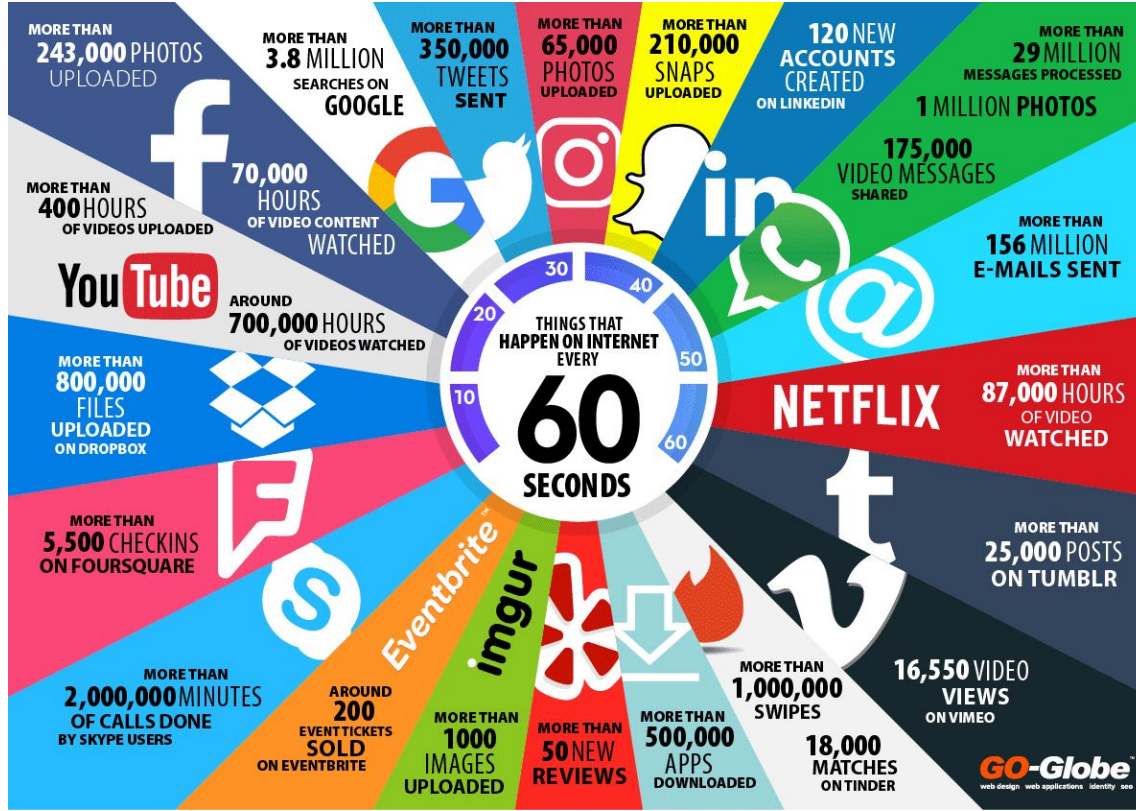
2.5 Exabytes of data generated by day

+ 90% was generated in the last 2 years

+ 90% are unstructured data

How to extract value from this new scenario?
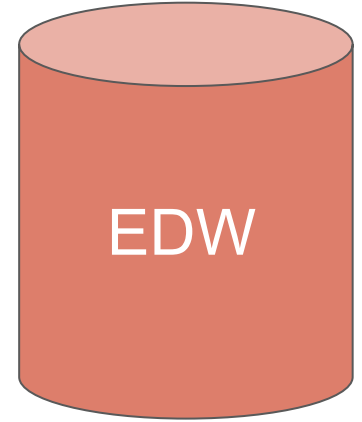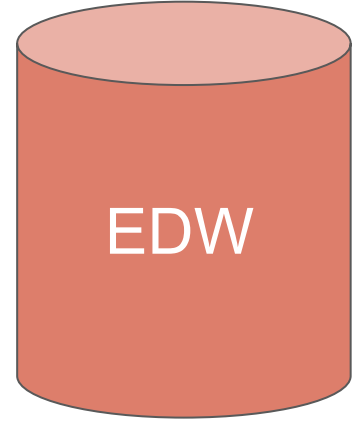
![Propus data science logo]

VOLUME **VARIETY** **VELOCITY**

Traditional Technologies
- scale vertically
- worker with structured data
- fixed and pre-structured data model

or

New Technologies
- scale horizontally
- worker with structured and unstructured data
- without previous modeling
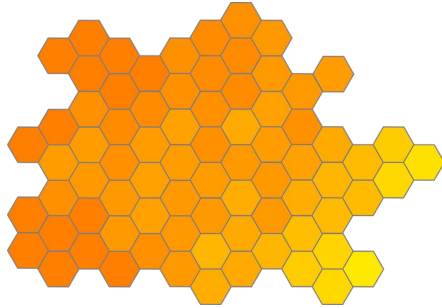
Distributed Storage

NoSQL / Advanced SQL

Distributed Processing

MapReduce / MPP

Why Open Source?

THE INNOVATION ADVANTAGE

INNOVATION

OPEN COMMUNITY

PROPRIETARY HADOOP

TIME

MAXIMUM COMMUNITY INNOVATION

100% Open Source Connected Data Platforms
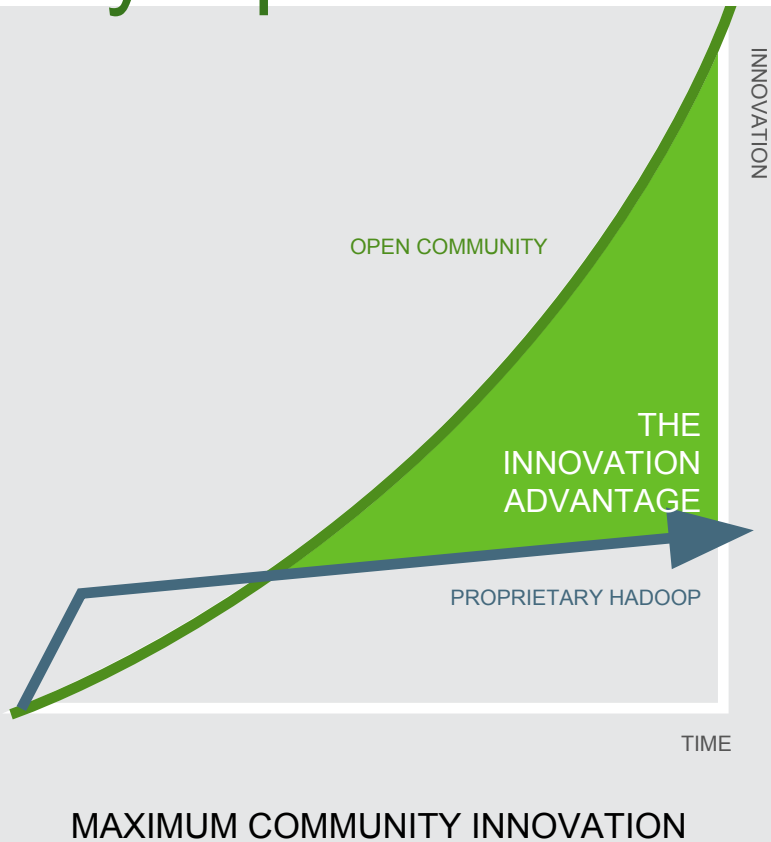
Eliminates Risk

of vendor lock-in by delivering 100% Apache open source technology

Maximizes Community Innovation

with hundreds of developers across hundreds of companies

Integrates Seamlessly

through committed co-engineering partnerships with other leading technologies

Hortonworks®

Data Lake

Corporate Data Lake

BIG DATA LANDSCAPE 2017

Propus data science

INFRASTRUCTURE

HADOOP ON-PREMISE
cloudera · Hortonworks · MAPR · IBM InfoSphere · Pivotal · bluedata · jethro

HADOOP IN THE CLOUD
amazon web services · Microsoft Azure · Google Cloud Platform · databricks · IBM InfoSphere BigInsights · TREASURE DATA · GridGain · METAMARKETS · confluent · Qubole · altiscale · CAZENA · CenturyLink

STREAMING / IN-MEMORY
amazon web services · striim · DataTorrent · dataArtisans · ORACLE · hazelcast · TERRACOTTA

NOSQL DATABASES
Google Cloud Platform · ORACLE · Amazon DynamoDB · Microsoft Azure · mongoDB · DATASTAX · AEROSPIKE · Couchbase · redislabs · influxdata

NEWSQL DATABASES
SAP · Clustrix · Pivotal · NuoDB · IBM · ORACLE · Cockroach LABS · memsql · splice · VoltDB · citusdata · Trafodion · deepdb · paradigm4

GRAPH DBS
neo4j · IBM · OrientDB · InfiniteGraph · Objectivity

MPP DBS
TERADATA · VERTICA · Netezza · Actian · ORACLE · Pivotal · EXASOL · dremio

CLOUD EDW
amazon · Google Cloud Platform · Microsoft Azure · Pivotal · snowflake · Infoworks

DATA TRANSFORMATION
talend · pentaho · alteryx · TRIFACTA · tamr · Paxata · StreamSets · UNIFI

DATA INTEGRATION
informatica · MuleSoft · SnapLogic · TEALIUM · Segment · WAVEFRONT · enigma · alooma · podium data · ZALONI · xplenty · import.io · Stitch

DATA GOVERNANCE
informatica · IBM · skyhigh · collibra · DATADOG · unravel · splunk · Procore · pagerduty · Alation · Waterline DATA · Numerify

MGMT / MONITORING
amazon · New Relic. · APPDYNAMICS · actifio

STORAGE
amazon · Google Cloud Platform · Microsoft Azure · ALLUXIO · nimblestorage · Qumulo · COHO · panasas

CLUSTER SERVICES
amazon · kubernetes · docker · MESOSPHERE · CoreOS · pepperdata

APP DEV
Lightbend · nervana · Upwork · WorkFusion

CROWDSOURCING
amazon mechanical turk · rainforest · CrowdFlower

HARDWARE
Google TPU · ARM · Graphcore · MYTHIC · cerebras · NVIDIA. · Movidius · SCORTEX

ANALYTICS

DATA ANALYST PLATFORMS
Microsoft · pentaho · alteryx · Digital Reasoning · GUAVUS · AYASDI · MATTVIO · Datameer · Quid · ClearStory · OrigamiLogic · interana · Bottlenose · ARIMO · ENDOR · MODE

DATA SCIENCE PLATFORMS
IBM · KNIME · dataiku · DOMINO · yhat · rapidminer · CONTINUUM ANALYTICS · Alpine · Angoss · ALGORITHMIA

BI PLATFORMS
Microsoft · DOMO · amazon web services · looker · Wave Analytics · ARCADIA DATA · GoodData · SiSENSE

VISUALIZATION
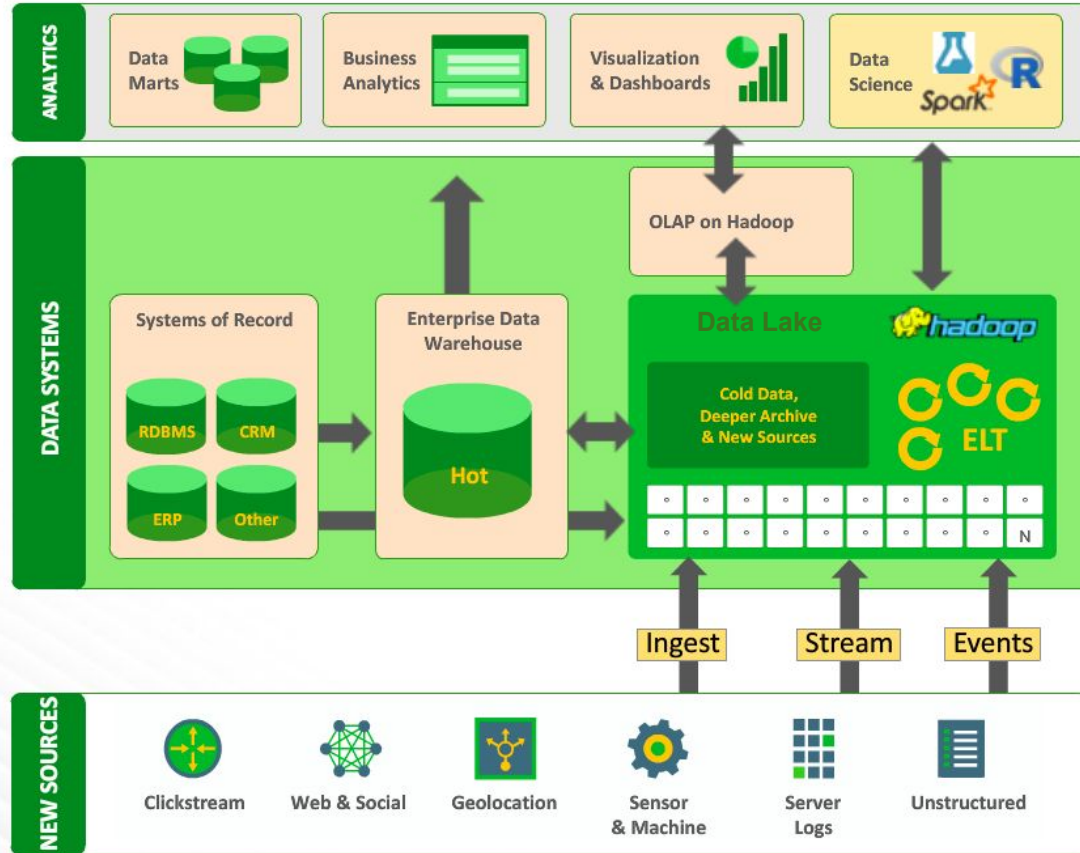tableau · SAP · Google Cloud Platform · Qlik · CHART.IO · Periscope · ZEPL · CHARTIO · plotly · CoMDATA

VERTICAL ANALYTICS
PREDIX · CAPE · UPTAKE · Orbital Insight · TACHYUS · Alluvium · datorama

STATISTICAL COMPUTING
SAS · SPSS · MATLAB

DATA SERVICES
Palantir · Six Sigma · OPERA · DATA SCIENCE · kaggle · Factual · EXL · DataKind · FF

MACHINE LEARNING
amazon machine learning · Google Cloud Platform · H2O.ai · DataRobot · ViSENZE · bonsai · DATAROBOT · context relevant · nutonian

HORIZONTAL AI
IBM Watson · Cortana. · Facet · Voyager.ai · vicarious · clarifai · Affectiva · CognitiveScale · x.ai · SENSE360 · Numenta · PETUUM · Data.XU · narq · OSARO · THE CURIOUS AI · SoundHound Inc.

SPEECH & NLP
Google Cloud Platform · twilio · amazon alexa · NarrativeScience · semantic machines · WellSaid Labs · MindMeld · ARRIA · NUANCE · Talki.O · cortical.io · maluuba · snips · xSeg

SEARCH
Autonomy · elastic · EXALEAD · ORACLE · ENDECA · ThoughtSpot · OpenX · Lucidworks · kibana · swiftype · MAANA · alphasense · Searchlink · SINEQUA

LOG ANALYTICS
splunk> · sumologic · loggly · kibana · logz.io

SOCIAL ANALYTICS
Hootsuite · sprinklr · NETBASE · DATASIFT · tracx · synthesio · simplereach · bitly · predata

WEB / MOBILE / COMMERCE ANALYTICS
Google Analytics · mixpanel · AMPLITUDE · sumAll · retention · Airtable · SIGOPT · granify · custora

APPLICATIONS – ENTERPRISE

SALES
einstein · CHORUS · INSIDESALES.COM · conversica · Gong · clari · AVISO · TACT · fusemachines · TROOPS

MARKETING - B2B
RADIUS · App Annie · EVERSTRING · Lattice · infer · MINTIGO · RK Reflektion · 6sense · tubular · Datafox · ENGAGIO

MARKETING - B2C
bloomreach · Zeta · blueyonder · [PERSADO] · kahuna · ACTIONIQ · BLUECORE · SAILTHRU · QUANTIFIND · mparticle · Amplero

CUSTOMER SERVICE
MEDALLIA · zendesk · CLARABRIDGE · Gainsight · ClickFox · NGDATA · DigitalGenius · appuri · AUTOMAT · frame.ai · msgai · INTERCOM

HUMAN CAPITAL
Hire-Vue · entelo · hiQ · GIGSTER · textio · RESTLESS BANDIT · Wade&Wendy · Clustree · Stella · pymetrics

LEGAL
RAVEL · Seal · Everlaw · JUDICATA · iumata · Clara · talla · butter.ai · KASIST.O

FINANCE
anaplan · zuora · SAP S/4 HANA · TRADESHIFT · ROSS · casetext

ENTERPRISE PRODUCTIVITY
slack · x.ai · ORACLE · Captricity · AppZen

BACK OFFICE AUTOMATION
HyperScience · talla

SECURITY
TANIUM · CYLANCE · StackPath · DARKTRACE · illumio · CODE42 · ThreatMetrix · DataGravity · VECTRA · CipherCloud · Qualatics Analytics · ANOMALI · sift science · SIGNIFYD · SentinelOne · BlueTalon · Recorded Future · feedzai · sqrrl · SOCURE · AREA 1 · FORTSCALE · Xaybeos · sparkcognition

APPLICATIONS – INDUSTRY

ADVERTISING
AppNexus · MediaMath · criteo · XAd · Integral · rocketfuel · OpenX · MOAT · theTradeDesk · drawbridge · Algorithmia · TAPAD · Data.XU · DYNAMIC YIELD · Yieldmo

EDUCATION
Knewton · Clever · declara · kidaptive · PANORAMA · know.re · gradescope

GOVERNMENT
Socrata · OPENGOV · mark43 · OpenDataSoft

FINANCE - LENDING
OnDeck · Affirm · Kreditech · AVANT · INSIKT · MoneyLion · cignifi · aire

FINANCE - INVESTING
Dataminr · INUMERI · TALA · TrueAccord · trooly · RavenPack

REAL ESTATE
Opendoor · VTS · reonomy · COMPSTAK

INSURANCE
Metromile · Lemonade · CYENCE · Shift Technology · Tractable

HEALTHCARE
FLATIRON · KYRUUS · HealthTap · METABIOTA · Ginger.io · BenevolentAI · COTA · zebra · AiCure · enlitic · imagia · Qventus · Kang Health · IMAGEN

LIFE SCIENCES
23andMe · Color · zymergen · TWIST · BIOSYSTEMS · Clear Labs · Phosphorus · OPTIMUS · nexar · two XAR · BAYLABS · Civil Maps · NIO

TRANSPORTATION
UBER · TESLA · nuTonomy · nauto · piloT.ai · OTTO · comma.ai · Terravion · netradyne · prospera

AGRICULTURE
FARMERS · instacart · FarmersEdge · FarmLogs · BLUE RIVER · mavrx · HowGood

COMMERCE
STITCH FIX · RetailNext · happr · BOXEVER · VERDIGRIS · Second Spectrum

OTHER
eHarmony · stem · rethink robotics · happr · celect · duetto · Unbabel · Jukedeck

CROSS-INFRASTRUCTURE/ANALYTICS

amazon web services · Google Cloud Platform · Microsoft · IBM · SAP · Hewlett Packard Enterprise · SAS · 1010data · vmware · TIBCO · TERADATA · ORACLE · NetApp

OPEN SOURCE

FRAMEWORK
Hadoop HDFS · Hadoop MapReduce · Flink · YARN · TEZ · MESOS · CDAP · Spark

QUERY / DATA FLOW
Spark SQL · presto · HIVE · SLAMDATA · APACHE DRILL · Google Cloud Dataflow

DATA ACCESS
cassandra · nifi · mongoDB · APEX · sqoop · CouchDB · OpenTSDB · riak · HBASE · Spanner · accumulo

COORDINATION
talend · Apache Zookeeper · Apache Ambari

STREAMING
Spark · Flink · beam · kafka · druid · STORM

STAT TOOLS
python · R · ScalaLab · NumPy · SciPy

AI / MACHINE LEARNING / DEEP LEARNING
TensorFlow · theano · Caffe · CNTK · DMTK · MXNet · leen · Keras · PaddlePaddle · Apache SINGA · OpenNLP · neon · FeatureFu · mxnet · VELES · WEKA · Chainer · DIMSUM · DSSTNE · mllib · DL4J · Aerosolve · minibatch

SEARCH
elasticsearch · Solr · Lucene

LOG ANALYSIS
elasticsearch · kibana · Rodeo · logstash

VISUALIZATION
BEAKER · Jupyter · Zeppelin

COLLABORATION
ANACONDA

SECURITY
Apache Ranger · KNOX · Sentry

DATA SOURCES & APIS

HEALTH
JAWBONE · VALIDIC · practicefusion · fitbit · GARMIN · Human API · kinsa

IOT
GE Digital · UPTAKE · ThingWorx · helium · samsara · AUGURY · estimote · Thinknum

FINANCIAL & ECONOMIC DATA
Bloomberg · THOMSON REUTERS · DOW JONES · S&P CAPITAL IQ · CBINSIGHTS · xignite · quandl · YODLEE · PREMISE · estimize · SECOND MEASURE · Eagle Alpha · StockTwits · PLAID · mattermark

AIR / SPACE / SEA
PLANET · Airware · Spire · SKYCATCH · AIRBOTICS · Descartes Labs · WINDWARD · TELLUSLABS · DroneDeploy · MarineTraffic

PEOPLE / ENTITIES
acxiom · Experian · EPSILON · InsideView · Crimson Hexagon · quantcast · BASIS · factual · SAFEGRAPH

LOCATION INTELLIGENCE
FOURSQUARE · MapAnything · Sense · PlaceIQ · esri · CARTO · Mapillary · STREETLINE

OTHER
Qualtrics · DATA.GOV · data.world · panjiva · enigma

DATA RESOURCES

INCUBATORS & SCHOOLS
PLURALSIGHT · galvanize · DataCamp · DataElite · INSIGHT · The Data Incubator · METIS

RESEARCH
facebook research · MIRI · OpenAI · CSAIL · ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE
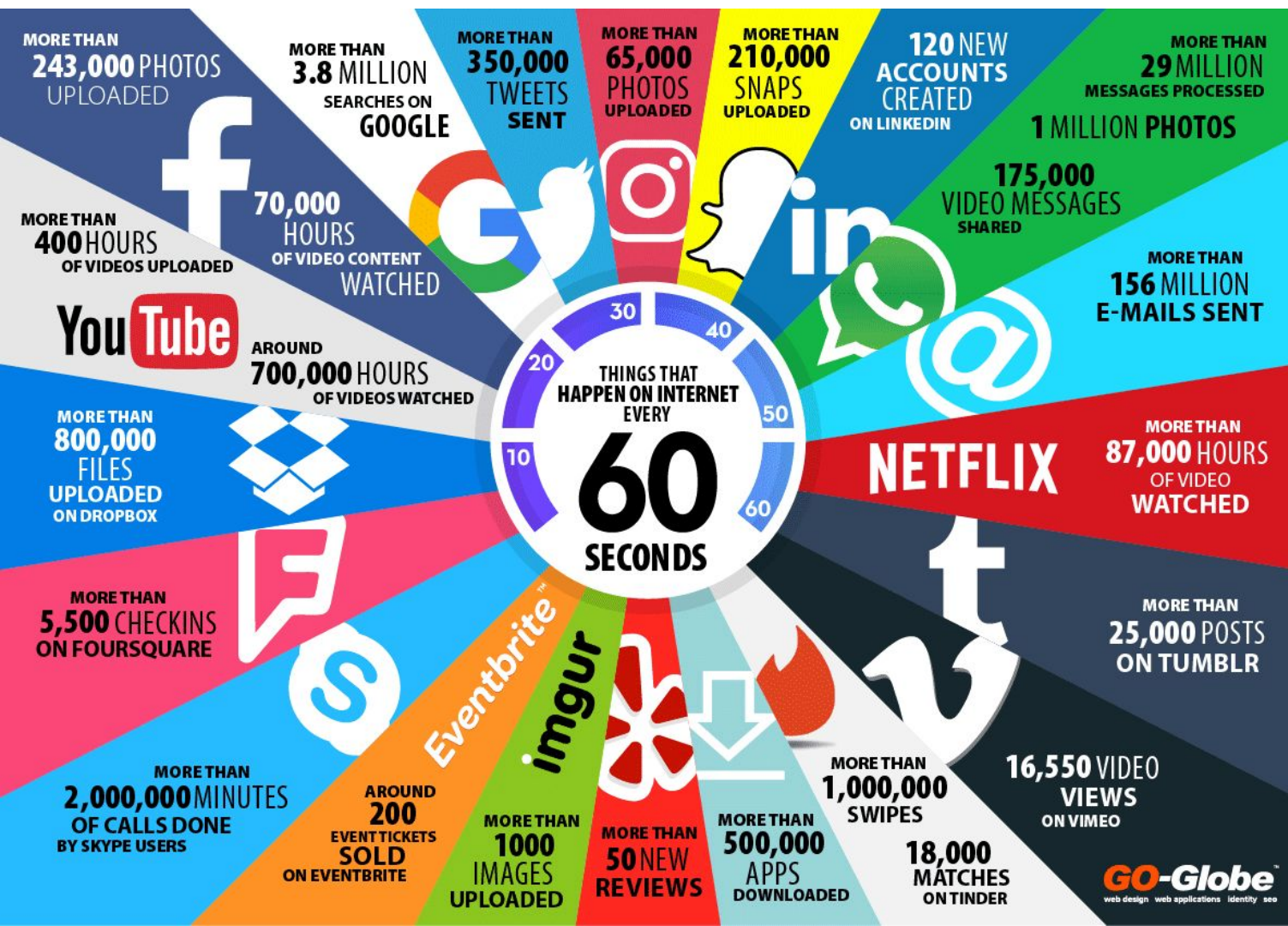
Open technologies can make open or proprietary products

Open Source Benefits

- No vendor lock-in

- Facilitates collaboration

- Reduced initial investment

- Worldwide community

- Security and flexible solutions

THINGS THAT HAPPEN ON INTERNET EVERY 60 SECONDS

MORE THAN 243,000 PHOTOS UPLOADED

MORE THAN 3.8 MILLION SEARCHES ON GOOGLE

MORE THAN 350,000 TWEETS SENT

MORE THAN 65,000 PHOTOS UPLOADED

MORE THAN 210,000 SNAPS UPLOADED

120 NEW ACCOUNTS CREATED ON LINKEDIN

MORE THAN 29 MILLION MESSAGES PROCESSED

1 MILLION PHOTOS

175,000 VIDEO MESSAGES SHARED

MORE THAN 400 HOURS OF VIDEOS UPLOADED

70,000 HOURS OF VIDEO CONTENT WATCHED

MORE THAN 156 MILLION E-MAILS SENT

AROUND 700,000 HOURS OF VIDEOS WATCHED

MORE THAN 800,000 FILES UPLOADED ON DROPBOX

NETFLIX

MORE THAN 87,000 HOURS OF VIDEO WATCHED

MORE THAN 5,500 CHECKINS ON FOURSQUARE

MORE THAN 25,000 POSTS ON TUMBLR

MORE THAN 2,000,000 MINUTES OF CALLS DONE BY SKYPE USERS

AROUND 200 EVENT TICKETS SOLD ON EVENTBRITE

MORE THAN 1000 IMAGES UPLOADED

MORE THAN 50 NEW REVIEWS

MORE THAN 500,000 APPS DOWNLOADED

MORE THAN 1,000,000 SWIPES

16,550 VIDEO VIEWS ON VIMEO

18,000 MATCHES ON TINDER

GO-Globe
web design web applications identity seo

Propus
data science

Data Privacy?

Open Data

- **Open data** is the idea that some data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.

- **Open data** may include non-textual material such as maps, genomes, connectomes, chemical compounds, mathematical and scientific formulae, medical data and practice, bioscience and biodiversity

Open Data in Science

The concept of **open access to scientific data** was institutionally established with the formation of the World Data Center system, in preparation for the **International Geophysical Year of 1957–1958**. The International Council of Scientific Unions (now the International Council for Science) oversees several World Data Centres with the mandate to minimize the risk of data loss and to maximize data accessibility.

Open Data in Government

At the international level, the United Nations has an open data website that publishes statistical data from Member States and UN Agencies, and The World Bank published a range of statistical data relating to developing countries. The European Commission has created two portals for the **European Union: the EU Open Data Portal** which gives access to open data from the EU institutions, agencies and other bodies and the PublicData portal that provides datasets from local, regional and national public bodies across Europe.

Open Data Commons - https://opendatacommons.org/

OpenData Barometer - http://opendatabarometer.org/

The Open Data Barometer

A global measure of how governments are publishing and using open data for accountability, innovation and social impact.

OpenData Institute - https://theodi.org/

Portal Brasileiro de Dados Abertos - http://dados.gov.br/

Take care about the data provenance in your Data Lake!

Thank you

Carlos Eurico Pittas do Canto

eurico@propus.science

www.linkedin.com/in/carloseuricopittasdocanto/

www.propus.science

Fone: +55 51-3024-3568