Applying “Big Data” Analytics in the Insurance Sector

Discussion led by Lu Li and Rachael McNaughton

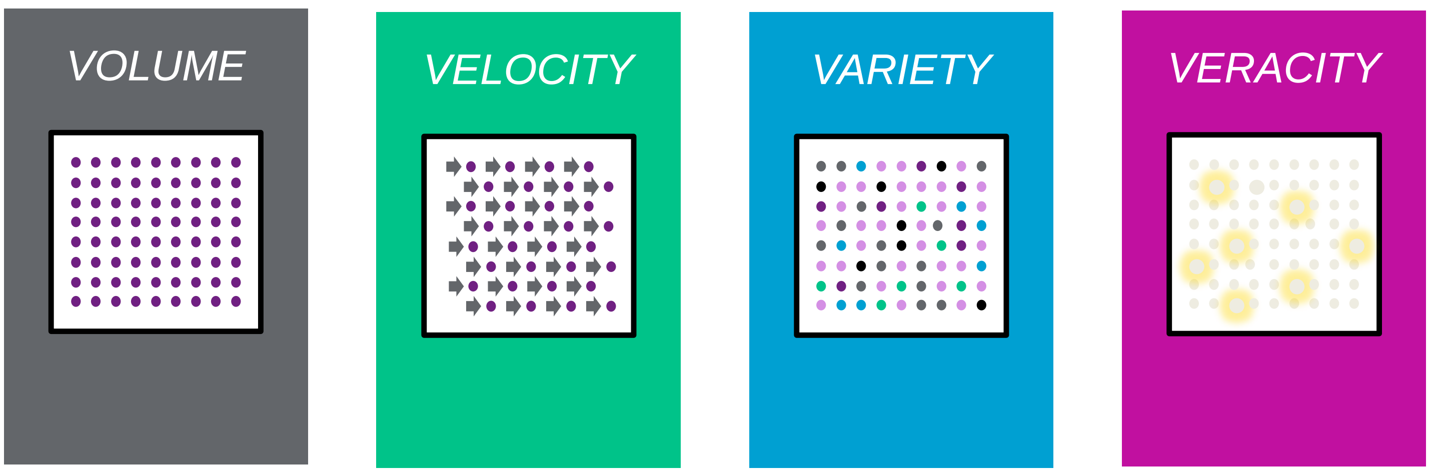
March 15, 2016

Willis Towers Watson

Defining big data

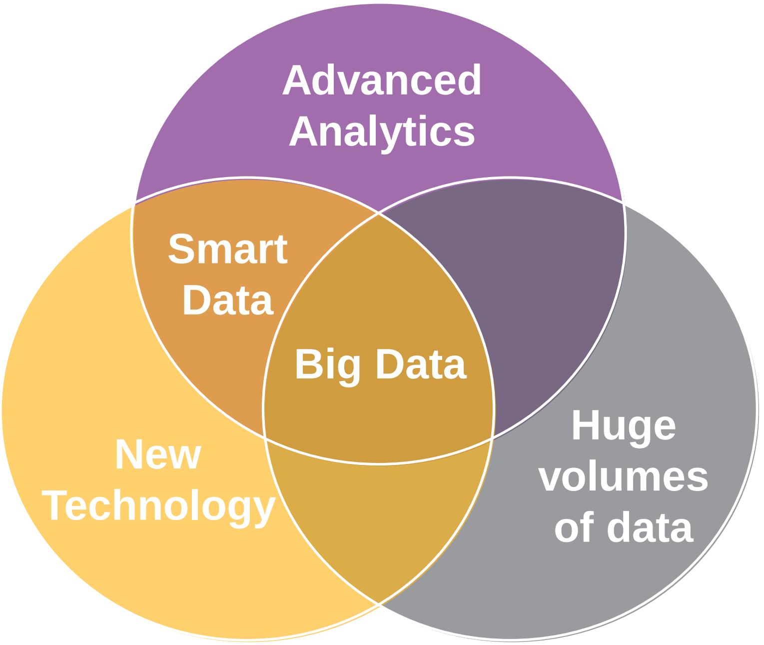
Big Data refers to both ***large volumes*** of data with ***high level of complexity*** and the analytical methods applied to them which require ***more advanced techniques and technologies*** in order to derive meaningful information and insights in ***real time***

HM GOVERNMENT HORIZON SCANNING PROGRAMME



How is big data affecting the insurance sector?

- Presently, genuine “big” data is less prevalent in the insurance sector than in some other industries
- But there is significant value in “smart” analytics via data set linkage and penetration
- External data enrichment, telematics, the IoT and clever insurance customer apps will shift the paradigm
- Smart analytics will then shift towards what data to keep and how to remove the noise



Should insurers take this seriously?

Contents

Contents

Cloud vs. on-premise

On-Premise (Cloudera, MapR, IBM)

- Pros
  - Full Control over the infrastructure
  - Data stays on site
  - Flexibility to build for specific purpose
  - Ability to install and configure specialist software
  - Long term planning – strong commitment
  - Cap Ex preferred
- Cons
  - High upfront investment
  - No on-demand scalability
  - Wasted capacity when not in use
  - Need in-house IT support
  - Ongoing maintenance required
  - Long implementation

Cloud (Azure, AWS, Bluemix, GCE)

- Pros
  - Low set-up costs
  - Scale elastically when needed
  - Support provided by cloud provider
  - Faster deployment
  - Reliability and fault-tolerance
  - Op Ex preferred
- Cons
  - No control over the infrastructure
  - Analytics tools from cloud provider
  - Data integration, security and privacy
  - Require internet connection
  - Availability disruptions and outages
  - Lease but not own

Hadoop infrastructure

- What is Hadoop?
- How to use it?
- Required Infrastructure:
  - Distributed Data Storage/Management: HDFS, HBase
  - Resource Management: YARN
  - Data Integration: Flume, Sqoop
  - Data Processing Engines: MapReduce, Spark
  - Other Applications:
    - Batch processing – MapReduce, Hive, Pig
    - Interactive data query – Impala, SparkSQL
    - Machine learning – Mahout, Spark (MLlib, MLI and ML Optimizer), Python (scikit-learn), C++ (mlpack), Java (Weka), C#/.net (accord)
    - Stream processing – Spark Streaming
    - Other languages – R, Perl, Julia, Scala

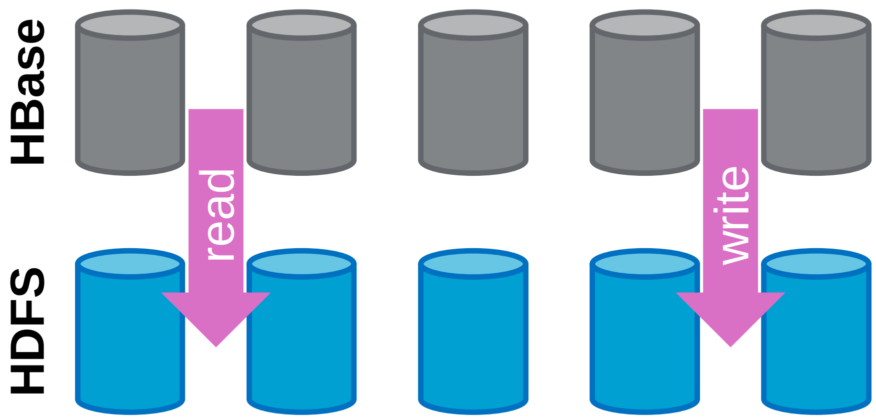
Hadoop infrastructure

HDFS

- Data storage layer of Hadoop
- Data is stored across multiple computers within a cluster
- Optimized for sequential access to a relatively small number of large files (e.g., > 100MB)
- A "write once read many" (WORM) file system

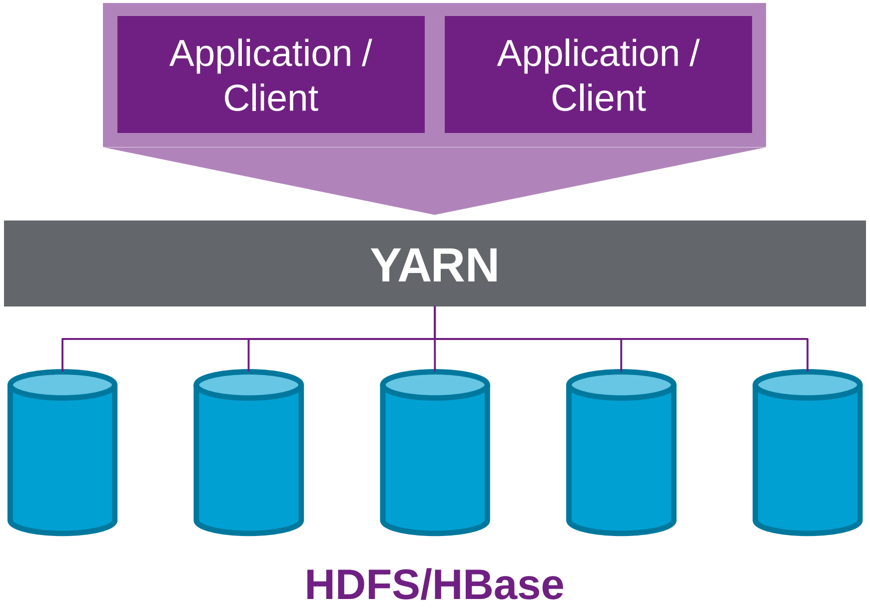
HBase

- Column-oriented database layer on top of HDFS (non-relational)
- Memory and CPU intensive
- Allows random read/write access to HDFS
- Adds transactional capabilities
  - Quick lookups, Inserts, Deletes, Updates



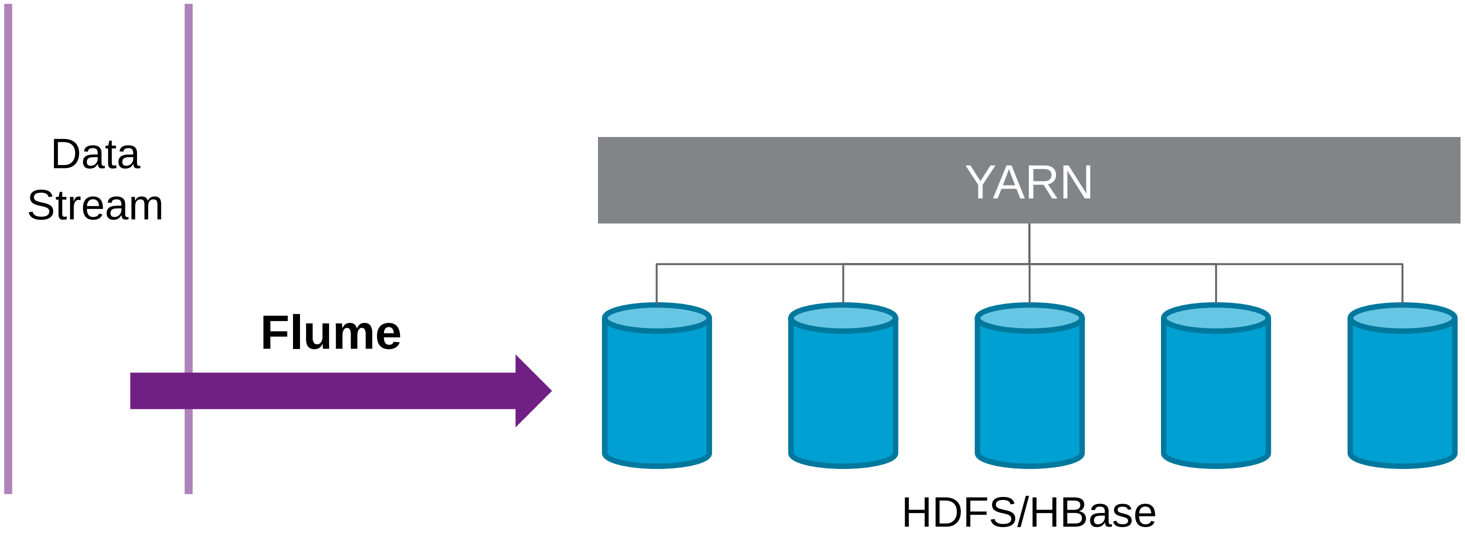
Hadoop infrastructure

- YARN (Yet Another Resource Negotiator)
  - Resource management system / distributed operating system
  - Negotiates resource requirements from the application with the distributed file systems it manages
  - Central resource manager + individual node manager
- Capacity vs. Fair Scheduler
  - Capacity scheduler allows you to setup queues to split resources
  - Fair scheduler allows you to split resources fairly across applications



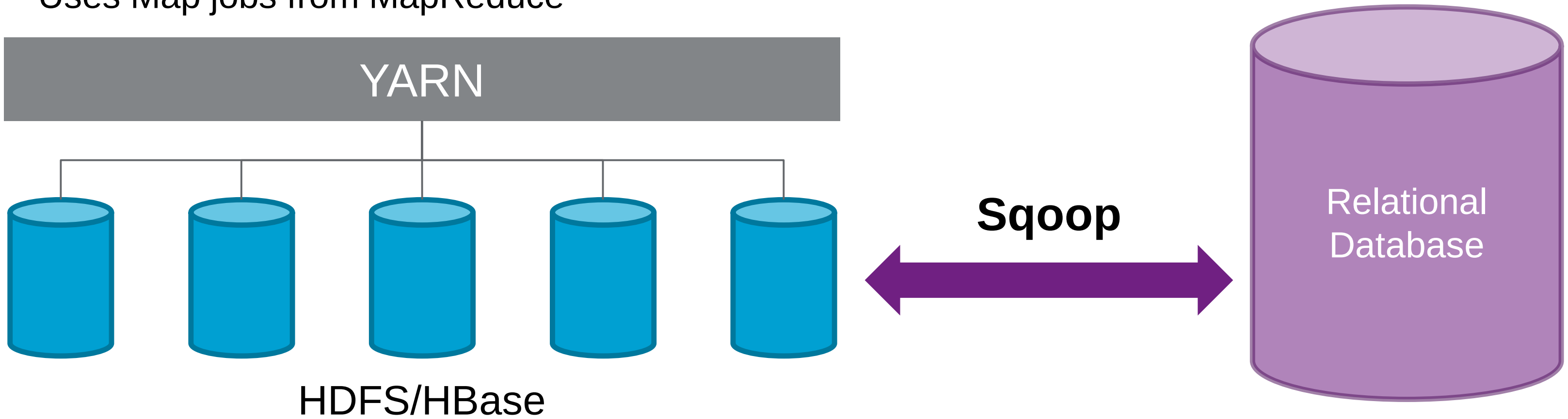
Hadoop infrastructure

- Flume
  - Designed for high-volume ingestion into Hadoop of event-based data, e.g., GPS / Application Logs / Digital Sensors
  - Distributed, scalable, reliable and manageable
  - Source (HTTP, JMS) → Channel (Memory, JDBC) → Sink (HDFS, Hbase)



Hadoop infrastructure

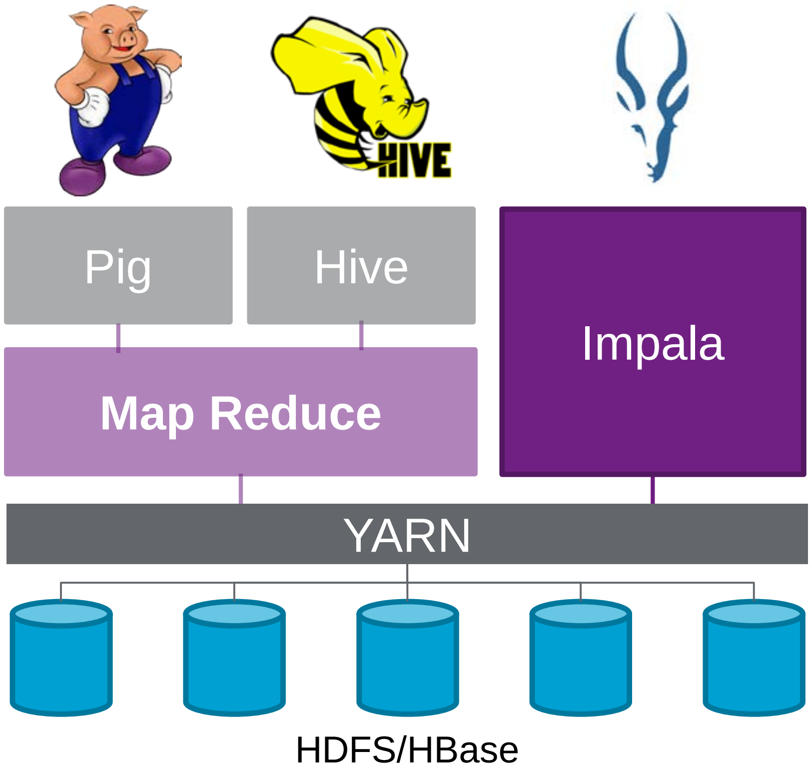
- Sqoop
  - Tool to transfer (import and export) data between Hadoop and relational databases and data warehouses
  - Works with relational databases such as Teradata, Netezza, Oracle, MySQL, Postgres, and HSQLDB
  - The dataset being transferred is sliced up into different partitions
  - Uses Map jobs from MapReduce



Data processing & analytics

MapReduce

- Paradigm for computation using a distributed environment (cluster)
- Written in Java but can be programmed using higher level abstractions such as Pig and Hive
- Impala is a specialized processing engine for interactive analysis
- Data locality, fault tolerance, linear scalability
- Each computation is split into two parts:
  - Map – data is split across multiple nodes and calculations are performed on each node independently
  - Reduce – Results are aggregated from all nodes according to the reduce function and the result returned to the client

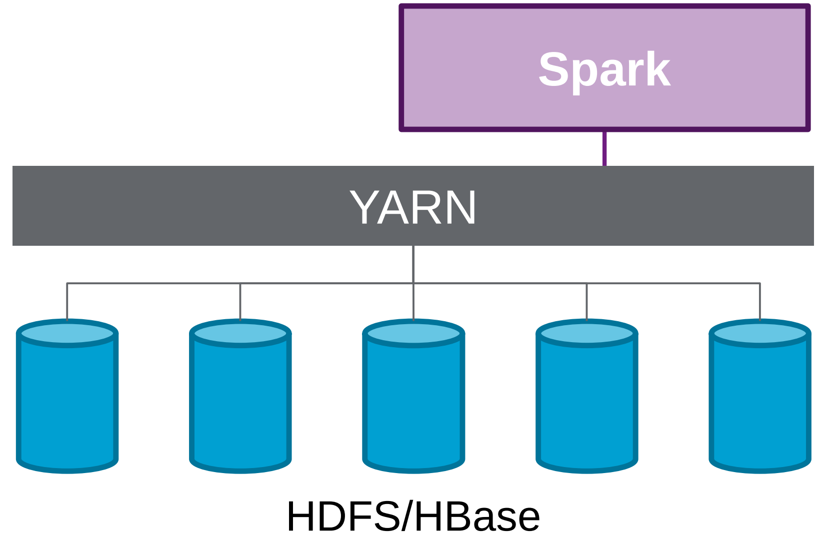


Data processing & analytics

Spark

- The new and up-coming data processing engine build on the same “map reduce” programming model as Hadoop MapReduce
- Improves efficiency (up to 100x faster) and usability (2-5x less code)
- All libraries work directly on RDDs
- “Bring the computation to the data”

- Computations are performed in-memory on individual cluster nodes
  - Reduces number of read/writes
  - Performs better with highly iterative algorithms than MapReduce
  - Theoretically limited by the amount of RAM available on each node
  - Built in machine learning library MLlib for analytics



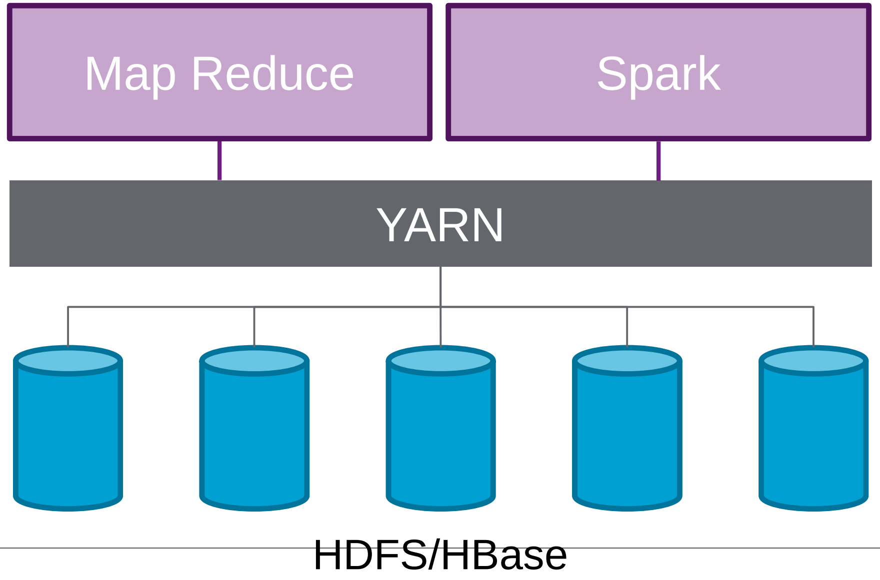
Data processing & analytics

MapReduce

- Java, Ruby, Perl, Python, PHP, R, C++
- On-Disk, Batch processing
- Runs job by dividing it into two types of tasks: map and reduce
- Need to integrate many disparate tools for advanced Big Data analytics: Hive, Pig, Flume, Sqoop, Mahout, Crunch

Spark

- Java, Scala, Python, R
- In-Memory, On-Disk, Batch, Interactive, Streaming (near real-time)
- The rapid in-memory processing of resilient distributed datasets (RDDs) is the core of Spark
- Spark Core, Spark Streaming, Spark SQL, Mllib, GraphX



Basic Infrastructure

Scalability!!

BI/MI Reporting

- Visualization, reporting and dashboard (Tableau, QlikView)
- 3rd party apps and services

Analytics

- Data mining (Hive, Pig, Perl)
- Descriptive (statistics, historical), predictive (forecasting, recommendation) and prescriptive (simulation, what-if) analytics (R, Python, Mahout)

Processing and Management

- Batch and In-Memory processing (MapReduce, Spark)
- Stream & Event (Spark, Flink)
- Resource management (Yarn, Zookeeper, Ozzie)

Data Storage

- Distributed File Systems (HDFS, GFS)
- Data Warehouses
- Databases (Hbase, Cassandra, MongoDB)

Data ETL

- Integration (Flume, Sqoop)

Data Sources

- Structured, Semi-structured, unstructured and streaming data

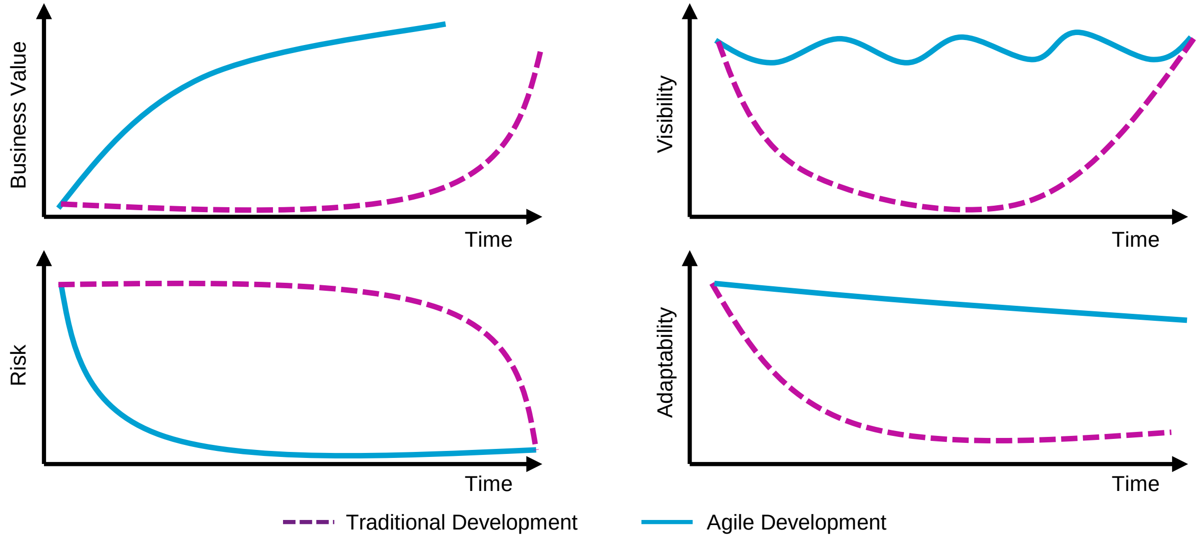
Contents

How we approach a big data project: The agile manifesto

Individuals and interactions over processes and tools
Working software over comprehensive documentation
Customer collaboration over contract negotiation
Responding to change over following a plan

Source: www.agilemanifesto.org

Why we believe it works? – Agile vs. Traditional approach



Source: "Agile vs Waterfall Visibility Ability to Change Business Value Risk (source: ADM) Waterfall Scrum 31--May--2012 effective agile. ex: www.slideshare.net728 × 514Search by image "

Scrum team organization

Stakeholders: Observe & advise

Scrum Master

- Knows scrum method very well
- Facilitates team and product owner by removing impediments
- NOT a project manager

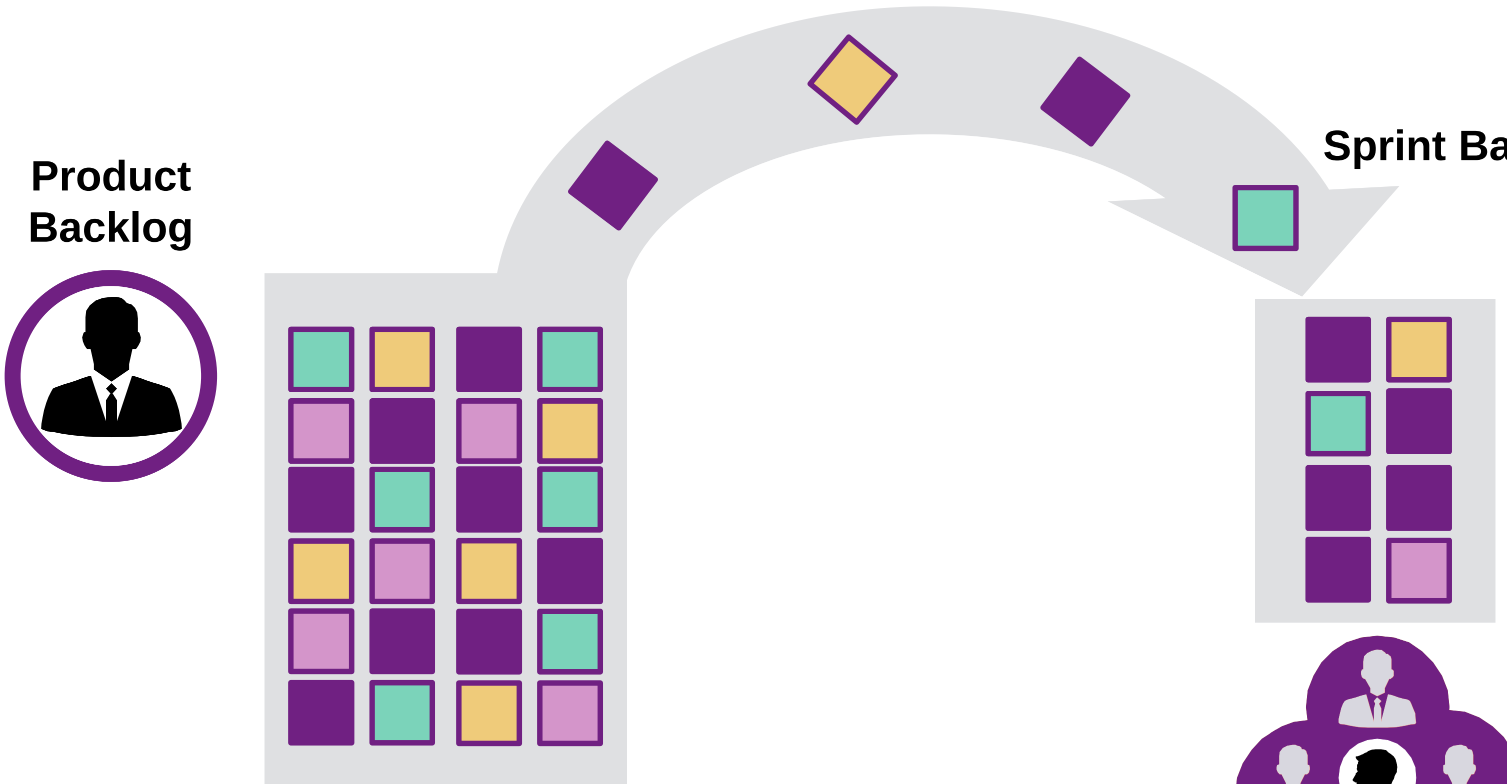
Product Owner

- Business lead
- Defines vision, requirements & priorities
- Accepts or reject team's deliverables
- Authority within the company

Team

- Estimates the work to be done
- Cross functional
- Builds the product as specified by owner
- Shared responsibility
- Assures quality of product
- Proactive
- Self organizing

Sprint planning



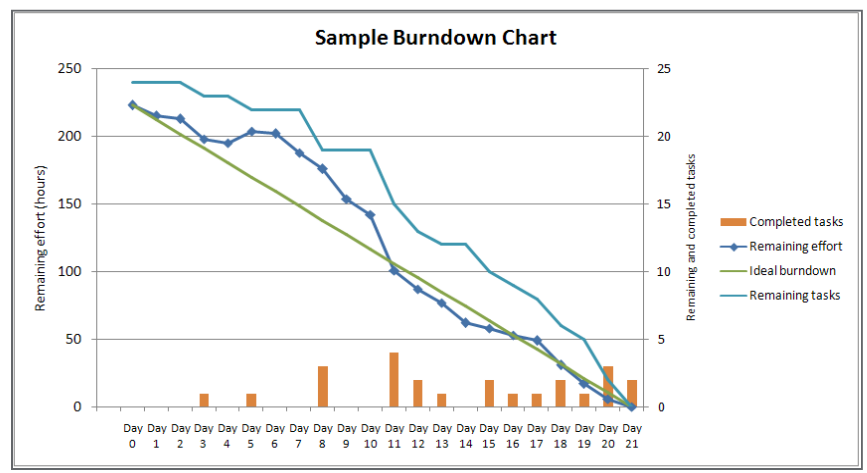
- Product owner presents candidate items
- Team estimates effort and budget for the items
- Product owner sets priorities for the Product Backlog
- Product owner sets Sprint Goal (one sentence summary)
- Team turns items into new Sprint Backlog

Work during the sprint

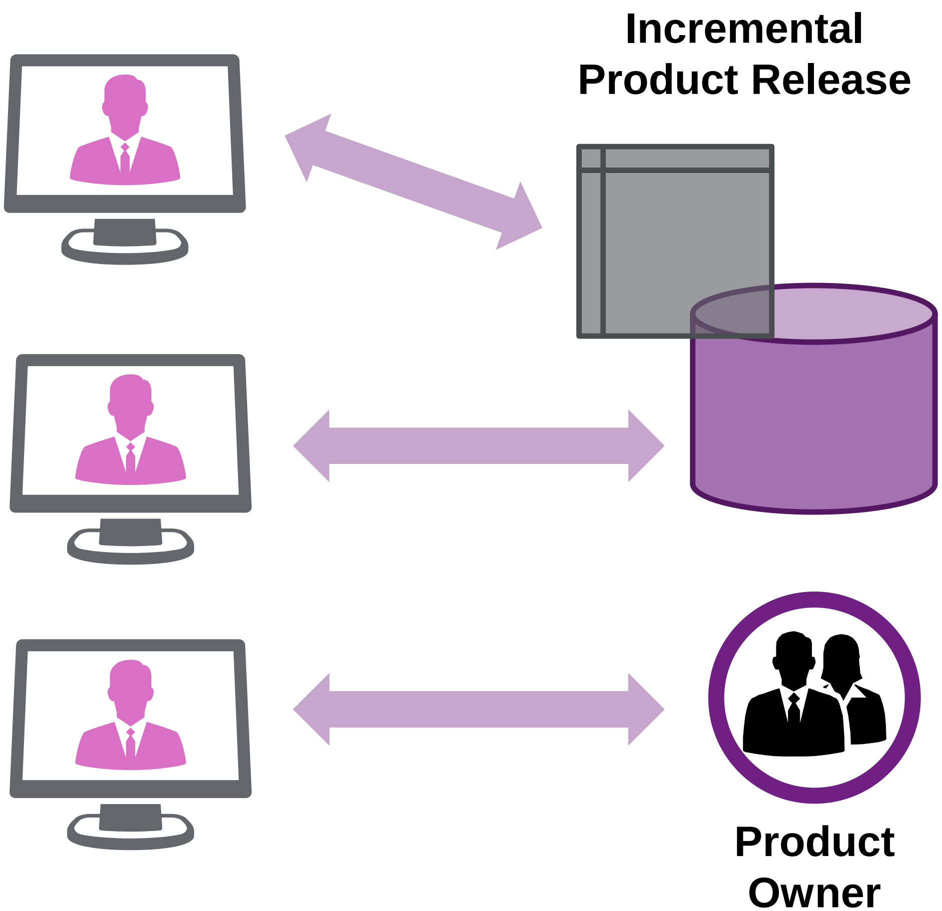
Daily Scrum:

- 15 mins same time every day
- Each talks about ...
  - What did you do yesterday?
  - What will you do today?
  - What's in your way?
- Team updates Sprint Backlog – Tasks are moved toward the Done column as they are completed

Total Remaining Work is tracked with a Burndown Chart



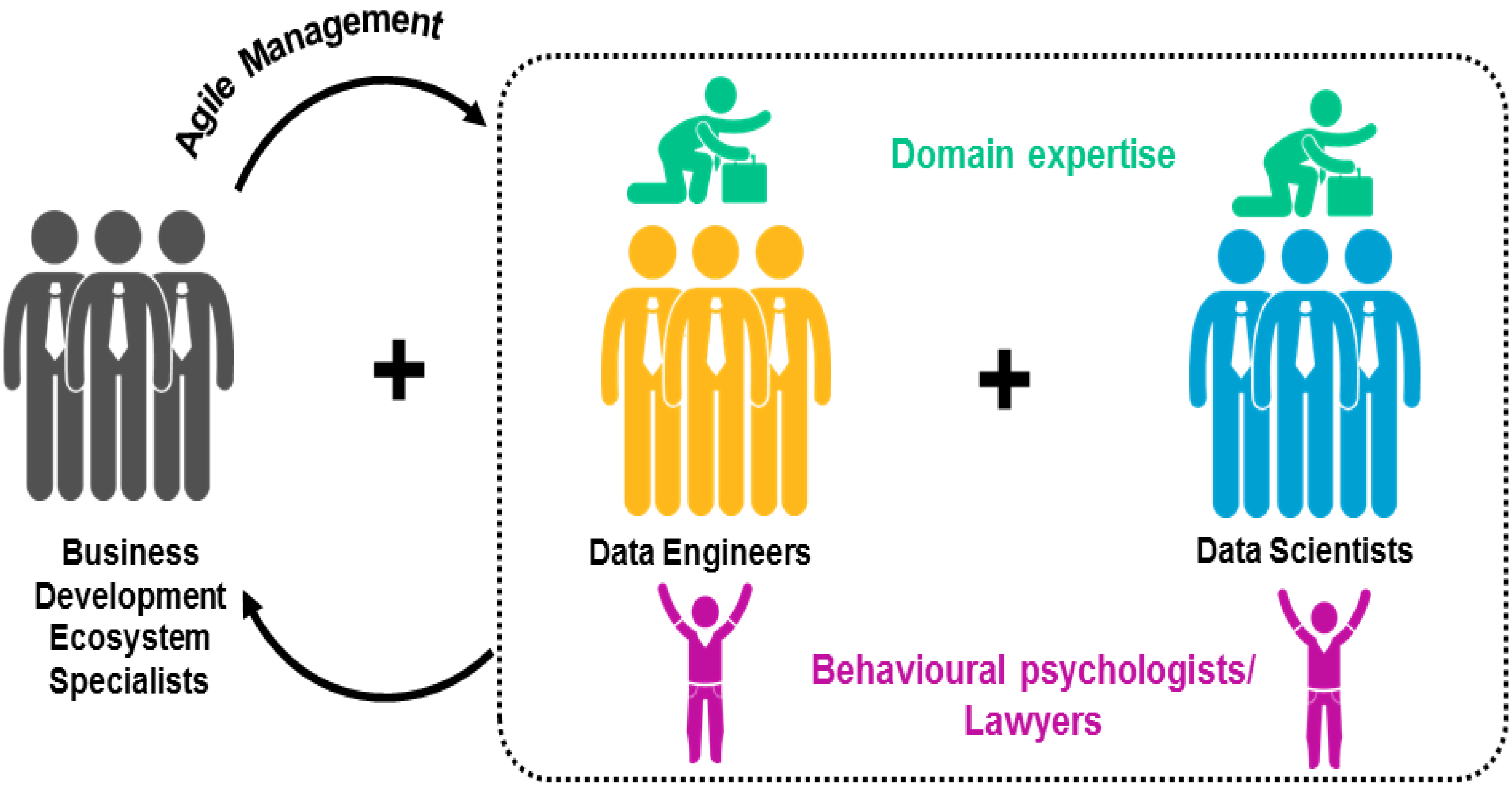
Sprint review



- ½ day at the end of the sprint
- Informal, informational
- Stakeholders provide feedback
- Agenda:
  - Product owner identifies what has been done
  - Team discusses what went well, what problem it ran into and those that were resolved
  - Team demonstrates what it has done
  - Product owner discusses the backlog as it stands
  - Entire group collaborates on what to do next

Contents

Transform Analytical Capabilities by a Team make-up



Data requirement

Access, Acquire & Collect

- Access internal data sources
- Acquire external data sources
- Collect open source data sets
- Future data collection (IoT)
- Centralized data repository

Quality Assessment

- Availability and granularity of raw data
- Functional information
- Timelines and accuracy
- Coherence
- Interpretability

Permission to use

Linkage

- Link data at customer, risk, transaction and constituency level
- Methods including
  * one-to-one
  * approximate comparison functions
  * rule-based, probabilistic classification
  * manual inspection (labor-intensive)

Identify Information Gap

- Due to data access, acquisition and collection
- Due to poor data quality
- Due to lack of data linkage

Data enrichment

Research

Continually identify new, innovative and potentially valuable data sources

Development

Routinely work with business units on proof of concepts

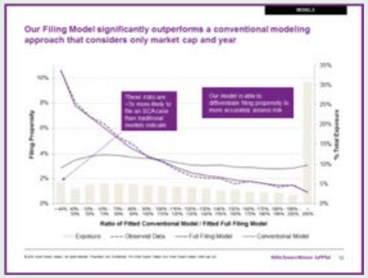
Value add

Consolidate, design and deploy these data products to enrich the analysis

- Geo-demographic and lifestyle (EASI, AGS)
- Weather (WeatherBank, NOAA)
- Crime
- Facebook personality profiles
- Wearables
- Credit/financial
- Property risks (e2Value, Explore, Opta)
- Auto (HLDI, Polk, CarFax, NHTSA)
- Commercial data (D&B, SIC and NAIC codes)
- Specialty lines (Advisen, AMA)

Advisen

- MSCAd contains SCA filings and associated settlement amounts for public companies traded on US exchanges
- WTW re-formatted and reconciled the case data and enhanced it:
  - Appended exposure data for public companies without historical cases, thereby enabling frequency models
  - Merged various company characteristics from securities and financial datasets



e2Value

- e2Value and WTW have partnered to create the Structure Insurance Score
- e2Value provides proprietary home and building characteristics, WTW provides predictive modeling expertise and sector knowledge
- Individual insurance companies provide historical experience

Modelling process

Data Design

- sources
- timeframe
- sampling
- fuzzy matching
- what is a record

Data Transformation & Visualization

- one-ways
- stats/common sense checking
- match rate
- standardization
- scaling
- normalization
- binarization
- encoding
- imputation

Data Exploration

- correlation
- linearity
- grouping
- clustering
- principle components
- independent components
- singular value decomposition
- factor analysis

Feature Generation & Selection

- high-order segmentation
- hierarchy separator
- sparse coding
- filter based
- permutation based
- formula based
- bag of words

Model Development

- supervised: GLM, LDA, SVM, SGD, NN, trees, Gaussian processes, naïve bayes, ensemble
- unsupervised: GMM, clustering, matrix factorization, NN, manifold learning
- **genetic equation search**

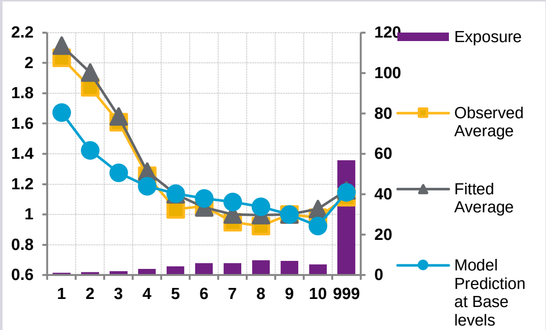
Model Validation

- holdout
- cross validation
- grid search
- quantifying the quality of predictions
- persistency
- validation curves

Example use cases

Motor pricing

- We found that clickstream behavioural variables form a strong predictor of motor risk, providing incremental value to well-established motor claims models.
  - Up to 1.8x increase in predictiveness for specific segments
  - Overall impact of similar scale as that of strong, well established rating variables

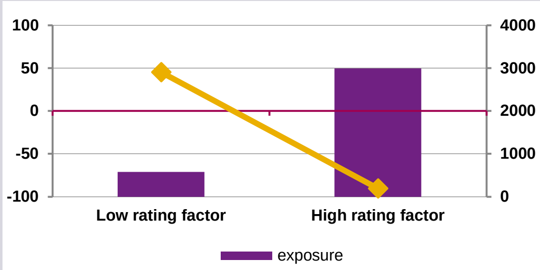


Cross-selling

- 200x higher propensity to cross-purchase between customer segments identified using our client's clickstream data, and past marketing campaigns variables (400+).
- This facilitated a different marketing approach (e.g. customer journey) based on a customer's web behaviour.

Underwriting

- We identified a new underwriting factor from open-source data which provided a strong predictor of ultimate contract profitability.
  - Rating factors drove a 15% variation in premium rate.



Q&A