BIG DATA TOOLS IN FOOD SAFETY

by: Dhea Fajriati Anas
Practice Case Big Data Tools

OUTLINE

01
Background

02
Big Data Workflow

03
Big Data Analytics with Azure

01
Background

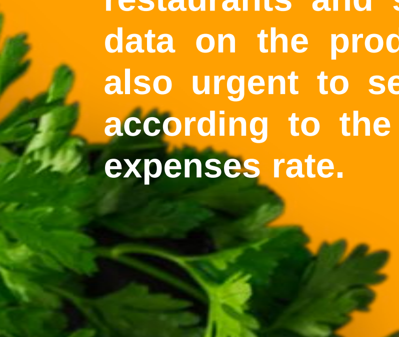
BACKGROUND

It is no wonder the food industry is one of the most vital industry segments for humanity. As consumers, we need our meal to be fresh, healthy and tasty. As stakeholders of the supply chain, we need full visibility and various info on customers' preferences, transportation status, restaurant prices, just to name a few.

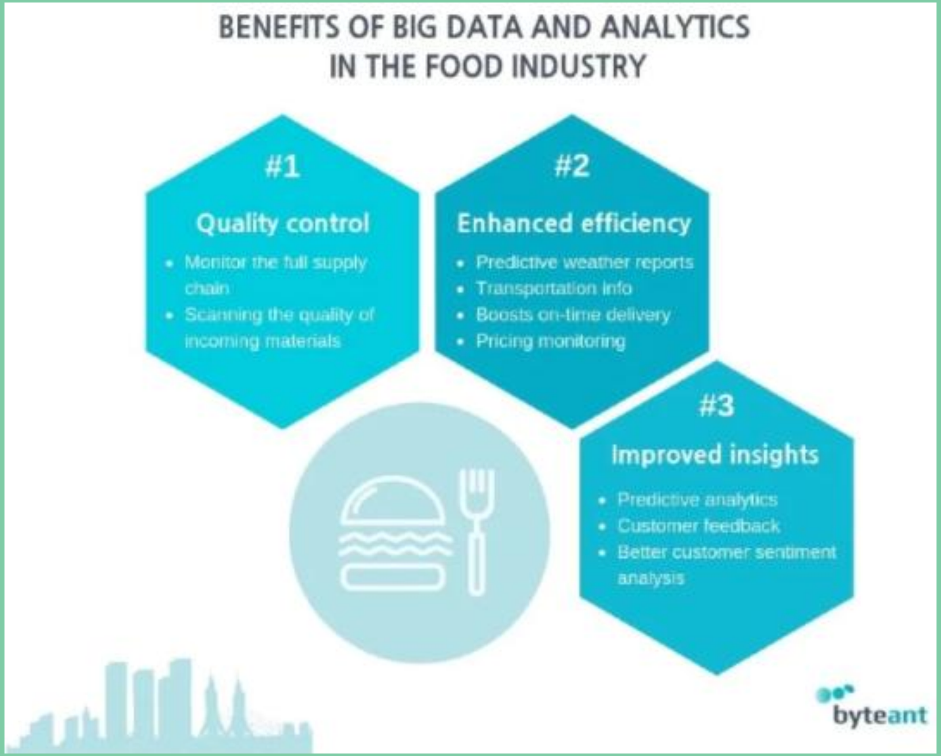
Every stakeholder, starting with farmers, shippers and retailers, ending with restaurants and shops, must have relevant data on the product and its condition. It's also urgent to see the full picture and act, according to the gathered data, as of high expenses rate.

FOOD & BEVERAGE BRANDS OF THE FUTURE
COFFEE & TEA
MEAT & DAIRY
ICE CREAM & FROYO
EGG & DAIRY ALTERNATIVES
BABY FOOD
NUTRITIONAL DRINKS
PROTEIN BARS
SNACK FOOD
INGREDIENTS & CONDIMENTS
PACKAGED MEALS
COCONUT WATER
WATER & FRUIT DRINKS
CANDY & GUM
PROBIOTIC DRINKS
ENERGY DRINKS
BLUE BOTTLE
NUMI
CALIFIA
TEABOX
PRE
good culture
IMPOSSIBLE
maia
yasso
CIAO BELLA
daiya
ripple
Hampton Creek
kite hill
melt
Heidi Ho
Clara Foods
soylent
mixo
HEALTH WARRIOR
exo
ALOHA
CBINSIGHTS
BONAFIDE
HILARY'S
Saffron Road
Urban Remedy
coco libre
VITA COCO
KEVITA
GoodBelly
humm
RUNA
GLUKOS
core
bai
zevia
hint
karma
suja
spindrift
VOSS
UNREAL
boomf
PRIMIZIE
CHIRPS
peeled
banana
nomva
HALFPOPS
nourish
RHYTHM
BRAD'S
dang
CHEF'S CUT
HALO
GOODNESS
two moms
SAHALE

(source: https://www.byteant.com/blog/how-big-data-is-boosting-food-industry-the-best-examples/)



BENEFITS OF BIG DATA



BENEFITS OF BIG DATA AND ANALYTICS IN THE FOOD INDUSTRY

#1 Quality control

- Monitor the full supply chain
- Scanning the quality of incoming materials

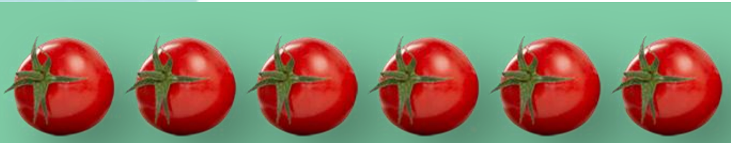
#2 Enhanced efficiency

- Predictive weather reports
- Transportation info
- Boosts on-time delivery
- Pricing monitoring

#3 Improved insights

- Predictive analytics
- Customer feedback
- Better customer sentiment analysis

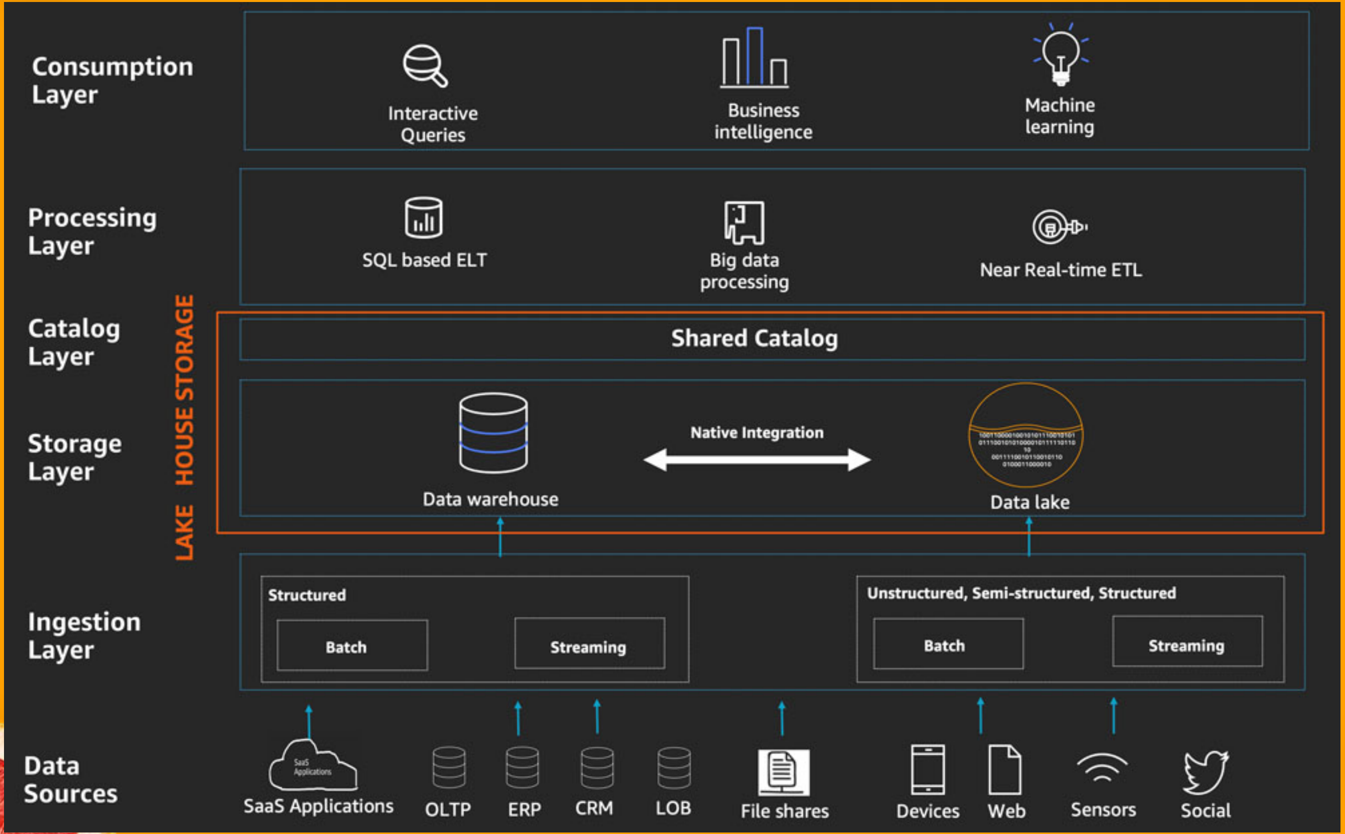
(source: https://www.byteant.com/blog/how-big-data-is-boosting-food-industry-the-best-examples/)



02

BIG DATA WORKFLOW

BIG DATA WORKFLOW



(source: https://aws.amazon.com/blogs/big-data/build-a-lake-house-architecture-on-aws/)

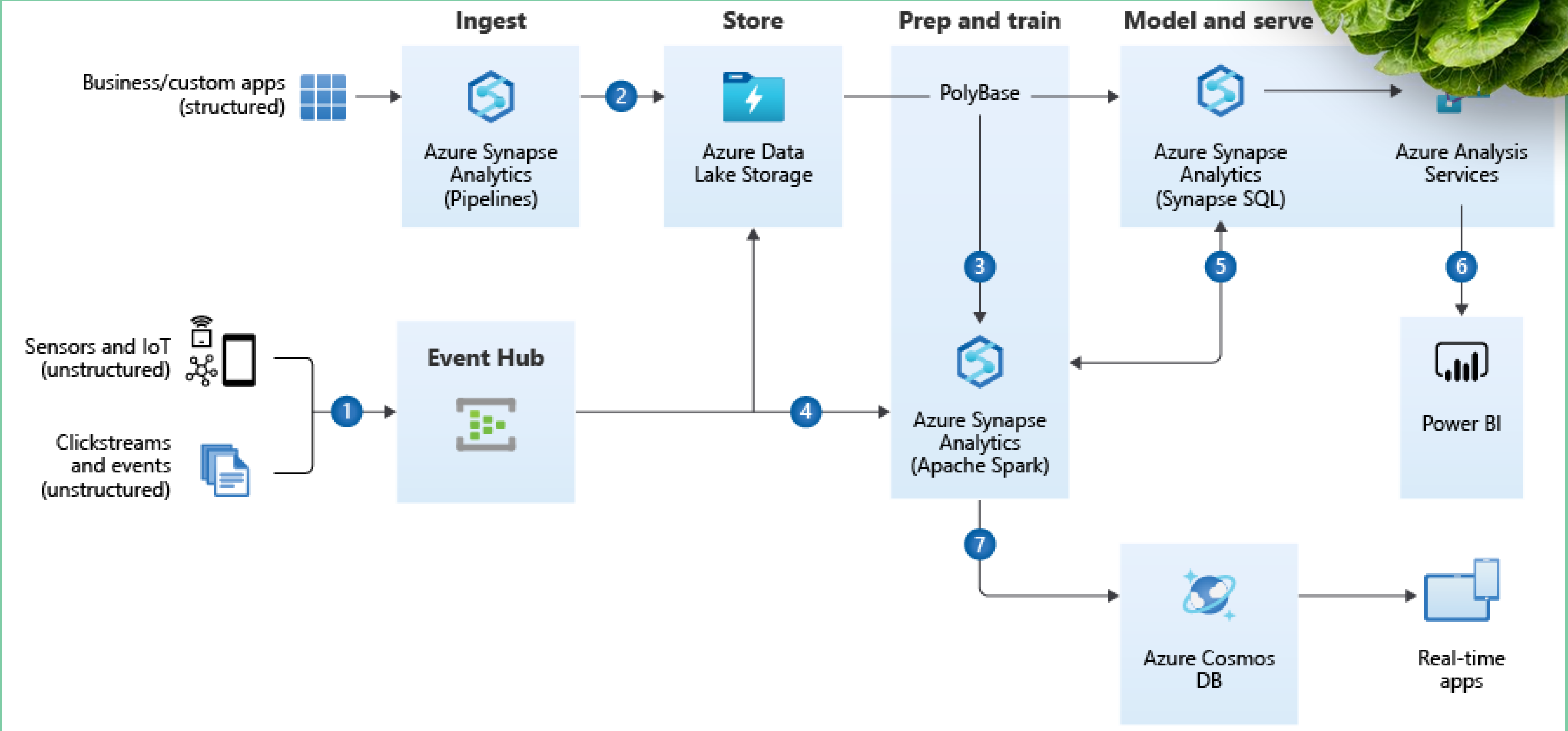
BIG DATA WORKFLOW

- **Data Source**
  The Lake House Architecture enables you to ingest and analyze data from a variety of sources.
- **Data Ingestion Layer**
  The ingestion layer in the Lake House Architecture is responsible for ingesting data into the Lake House storage layer. It provides the ability to connect to internal and external data sources over a variety of protocols.
- **Data Storage Layer**
  The data storage layer of the Lake House Architecture is responsible for providing durable, scalable, and cost-effective components to store and manage vast quantities of data. Data stored in a warehouse is typically sourced from highly structured internal and external sources. A data lake is the centralized data repository that stores all of an organization’s data. It supports storage of data in structured, semi-structured, and unstructured formats.
- **Catalog Layer**
  The catalog layer is responsible for storing business and technical metadata about datasets hosted in the Lake House storage layer.
- **Lake House Interface**
  In the Lake House Architecture, the data warehouse and data lake are natively integrated at the storage as well as common catalog layers to present unified a Lake House interface to processing and consumption layers.
- **Data Processing Layer**
  Components in the data processing layer of the Lake House Architecture are responsible for transforming data into a consumable state through data validation, cleanup, normalization, transformation, and enrichment.
- **Data Consumption Layer**
  The data consumption layer of the Lake house Architecture is responsible for providing scalable and performant components that use unified Lake House interfaces to access all the data stored in Lake House storage and all the metadata stored in the Lake House catalog

03

Big Data Analytics with Azure

BIG DATA ANALYTICS WITH AZURE



Ingest
Store
Prep and train
Model and serve
Business/custom apps (structured)
Azure Synapse Analytics (Pipelines)
2
Azure Data Lake Storage
PolyBase
Azure Synapse Analytics (Synapse SQL)
Azure Analysis Services
3
5
6
Sensors and IoT (unstructured)
Clickstreams and events (unstructured)
1
Event Hub
4
Azure Synapse Analytics (Apache Spark)
Power BI
7
Azure Cosmos DB
Real-time apps

(source: https://docs.microsoft.com/en-us/azure/architecture/solution-ideas/articles/advanced-analytics-on-big-data)

BIG DATA ANALYTICS WITH AZURE

DATA FLOW

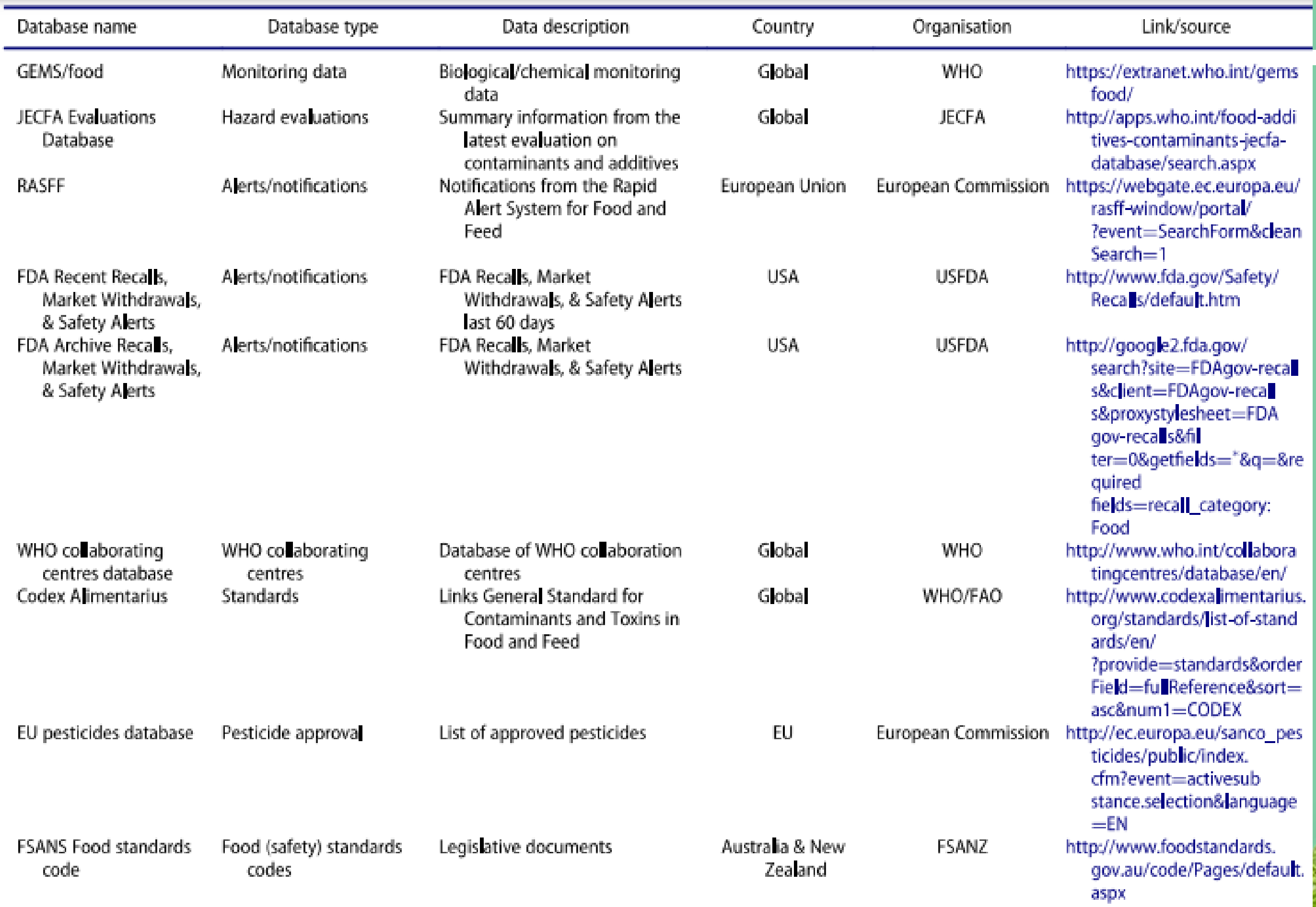
1. Structured, unstructured, and semi-structured data (logs, files, and media) using Synapse Pipelines to Azure Data Lake Storage.
2. Use Apache Spark pools to clean and transform the structureless datasets and combine them with structured data from operational databases or data warehouses.
3. Use scalable machine learning/deep learning techniques, to derive deeper insights from this data using Python, Scala, or .NET, with notebook experiences in Apache Spark pool.
4. Apply Apache Spark pool and Synapse Pipelines in Azure Synapse Analytics to access and move data at scale.
5. Query and report on data in Power BI
6. Take the insights from Apache Spark pools to Cosmos DB to make them accessible through web and mobile apps.

BIG DATA ANALYTICS WITH AZURE

COMPONENTS

1. Azure synapse analytics is the fast, flexible, and trusted cloud data warehouse that lets you scale, compute, and store elastically and independently, with a massively parallel processing architecture.
2. Synapse pipelines documentation allows you to create, schedule, and orchestrate your etl/elt workflows.
3. Azure blob storage is a massively scalable object storage for any type of unstructured data-images, videos, audio, documents, and more-easily and cost-effectively.
4. Azure synapse analytics spark pools is a fast, easy, and collaborative apache spark-based analytics platform.
5. Azure Cosmos DB is a globally distributed, multi-model database service. Learn how to replicate your data across any number of azure regions and scale your throughput independent from your storage.
6. Azure analysis services is an enterprise grade analytics as a service that lets you govern, deploy, test, and deliver your bi solution with confidence.
7. Power BI is a suite of business analytics tools that deliver insights throughout your organization. Connect to hundreds of data sources, simplify data prep, and drive unplanned analysis. Produce beautiful reports, then publish them for your organization to consume on the web and across mobile devices.

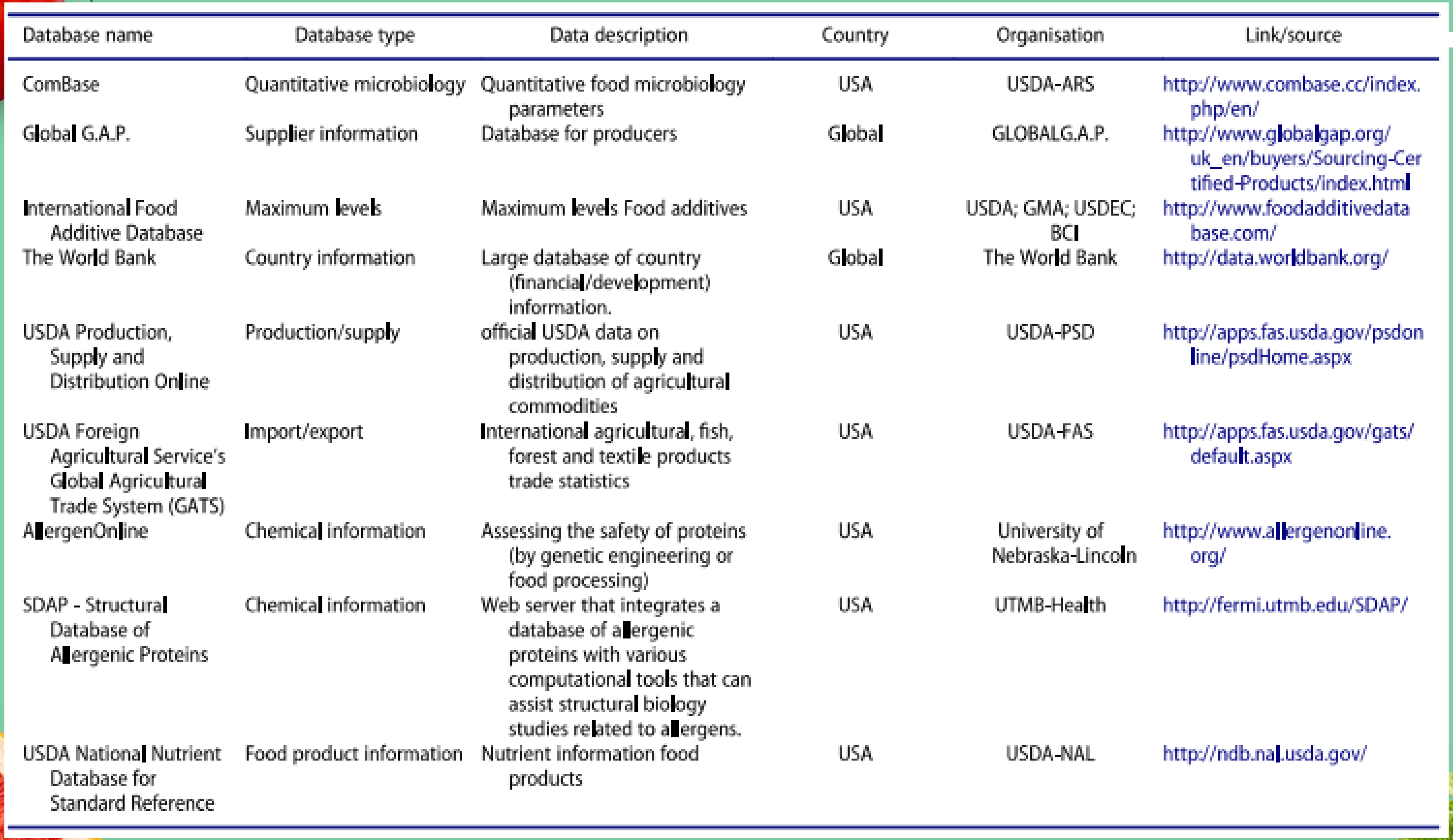
DATA COLLECTION IN FOOD SAFETY



Database name	Database type	Data description	Country	Organisation	Link/source
GEMS/food	Monitoring data	Biological/chemical monitoring data	Global	WHO	https://extranet.who.int/gemsfood/
JECFA Evaluations Database	Hazard evaluations	Summary information from the latest evaluation on contaminants and additives	Global	JECFA	http://apps.who.int/food-additives-contaminants-jecfa-database/search.aspx
RASFF	Alerts/notifications	Notifications from the Rapid Alert System for Food and Feed	European Union	European Commission	https://webgate.ec.europa.eu/rasff-window/portal/?event=SearchForm&cleanSearch=1
FDA Recent Recalls, Market Withdrawals, & Safety Alerts	Alerts/notifications	FDA Recalls, Market Withdrawals, & Safety Alerts last 60 days	USA	USFDA	http://www.fda.gov/Safety/Recalls/default.htm
FDA Archive Recalls, Market Withdrawals, & Safety Alerts	Alerts/notifications	FDA Recalls, Market Withdrawals, & Safety Alerts	USA	USFDA	http://google2.fda.gov/search?site=FDAgov-recalls&client=FDAgov-recalls&proxystylesheet=FDAgov-recalls&filter=0&getfields=*&q=&requiredfields=recall_category:Food
WHO collaborating centres database	WHO collaborating centres	Database of WHO collaboration centres	Global	WHO	http://www.who.int/collaboratingcentres/database/en/
Codex Alimentarius	Standards	Links General Standard for Contaminants and Toxins in Food and Feed	Global	WHO/FAO	http://www.codexalimentarius.org/standards/list-of-standards/en/?provide=standards&orderField=fullReference&sort=asc&num1=CODEX
EU pesticides database	Pesticide approval	List of approved pesticides	EU	European Commission	http://ec.europa.eu/sanco_pesticides/public/index.cfm?event=activesubstance.selection&language=EN
FSANS Food standards code	Food (safety) standards codes	Legislative documents	Australia & New Zealand	FSANZ	http://www.foodstandards.gov.au/code/Pages/default.aspx

(source: https://www.researchgate.net/publication/309740042_Big_data_in_food_safety_An_overview)

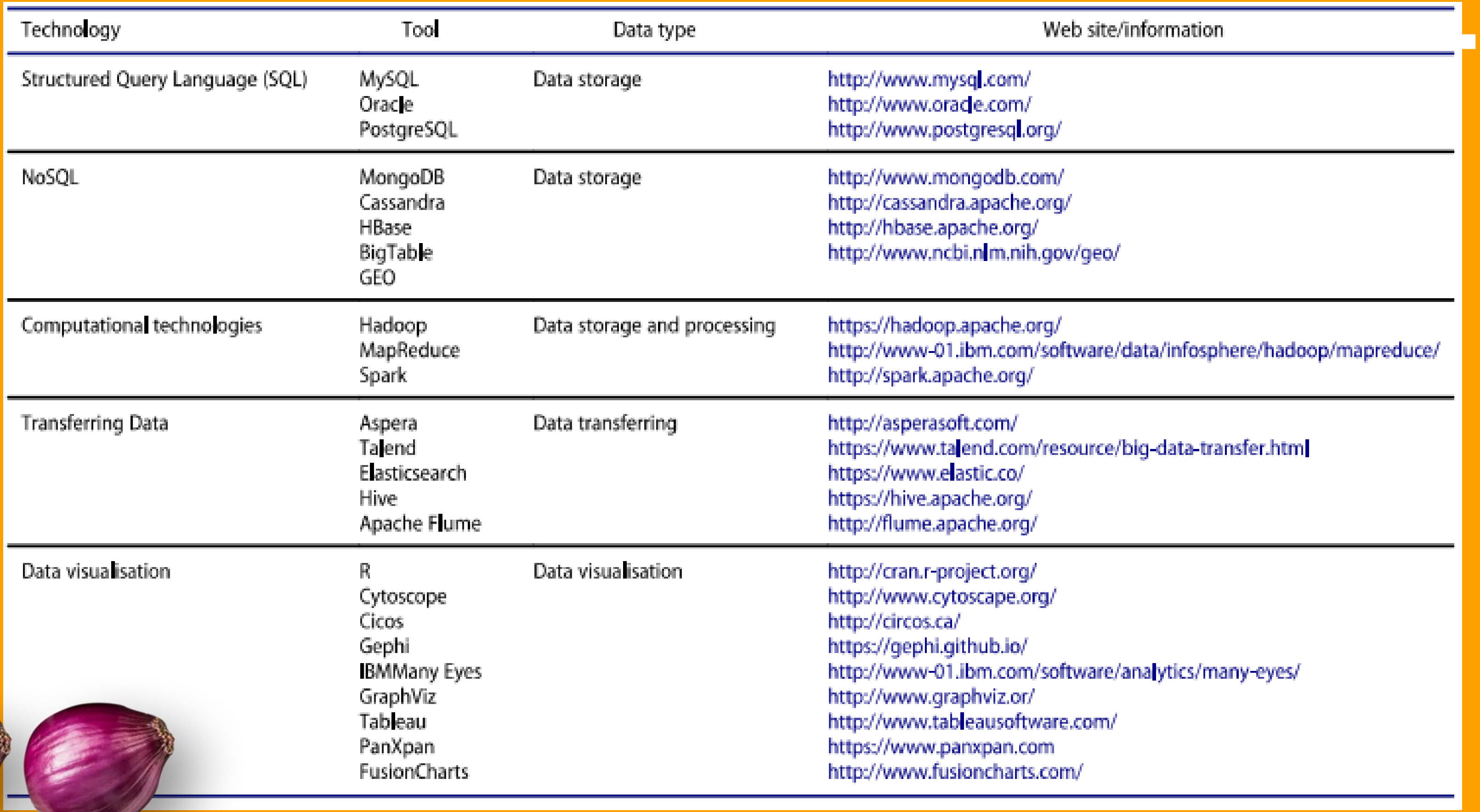
DATA COLLECTION IN FOOD SAFETY



Database name	Database type	Data description	Country	Organisation	Link/source
ComBase	Quantitative microbiology	Quantitative food microbiology parameters	USA	USDA-ARS	http://www.combase.cc/index.php/en/
Global G.A.P.	Supplier information	Database for producers	Global	GLOBALG.A.P.	http://www.globalgap.org/uk_en/buyers/Sourcing-Certified-Products/index.html
International Food Additive Database	Maximum levels	Maximum levels Food additives	USA	USDA; GMA; USDEC; BCI	http://www.foodadditivedatabase.com/
The World Bank	Country information	Large database of country (financial/development) information.	Global	The World Bank	http://data.worldbank.org/
USDA Production, Supply and Distribution Online	Production/supply	official USDA data on production, supply and distribution of agricultural commodities	USA	USDA-PSD	http://apps.fas.usda.gov/psdonline/psdHome.aspx
USDA Foreign Agricultural Service's Global Agricultural Trade System (GATS)	Import/export	International agricultural, fish, forest and textile products trade statistics	USA	USDA-FAS	http://apps.fas.usda.gov/gats/default.aspx
AllergenOnline	Chemical information	Assessing the safety of proteins (by genetic engineering or food processing)	USA	University of Nebraska-Lincoln	http://www.allergenonline.org/
SDAP - Structural Database of Allergenic Proteins	Chemical information	Web server that integrates a database of allergenic proteins with various computational tools that can assist structural biology studies related to allergens.	USA	UTMB-Health	http://fermi.utmb.edu/SDAP/
USDA National Nutrient Database for Standard Reference	Food product information	Nutrient information food products	USA	USDA-NAL	http://ndb.nal.usda.gov/

(source: https://www.researchgate.net/publication/309740042_Big_data_in_food_safety_An_overview)

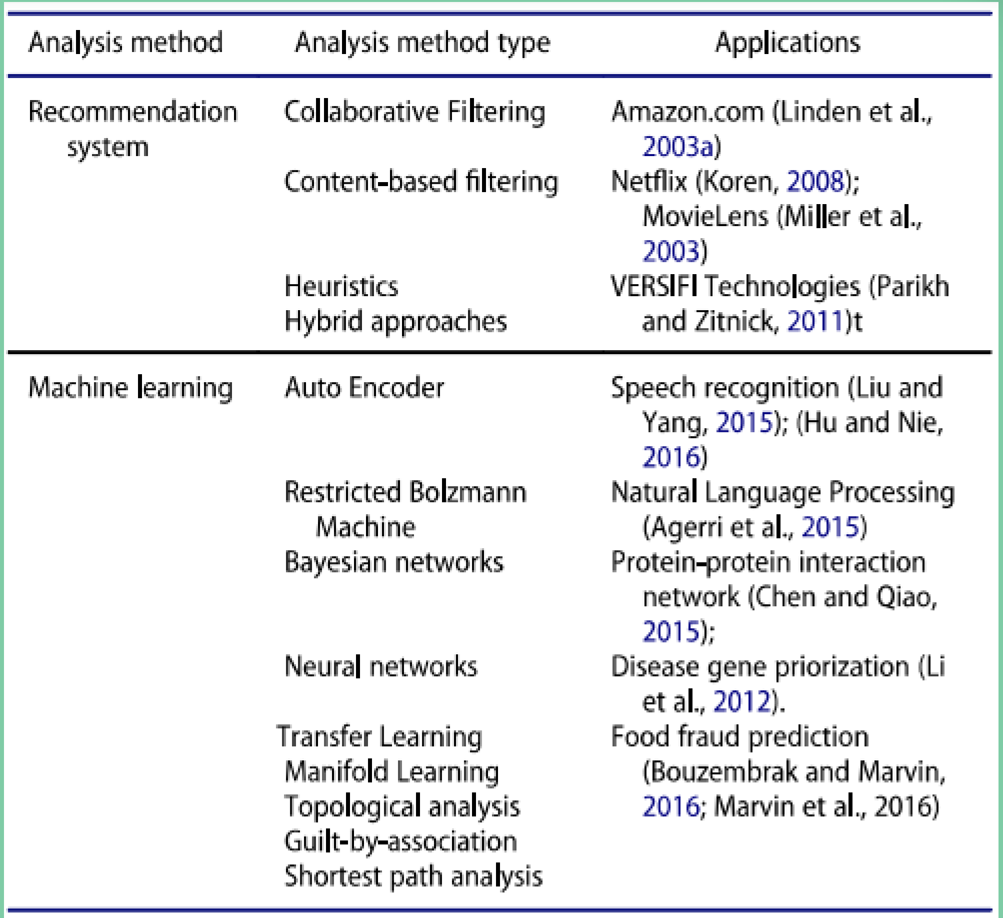
ANOTHER EXAMPLES OF DATA STORAGE, PROCESSING, AND VISUALIZATION



Technology	Tool	Data type	Web site/information
Structured Query Language (SQL)	MySQL Oracle PostgreSQL	Data storage	http://www.mysql.com/ http://www.oracle.com/ http://www.postgresql.org/
NoSQL	MongoDB Cassandra HBase BigTable GEO	Data storage	http://www.mongodb.com/ http://cassandra.apache.org/ http://hbase.apache.org/ http://www.ncbi.nlm.nih.gov/geo/
Computational technologies	Hadoop MapReduce Spark	Data storage and processing	https://hadoop.apache.org/ http://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/ http://spark.apache.org/
Transferring Data	Aspera Talend Elasticsearch Hive Apache Flume	Data transferring	http://asperasoft.com/ https://www.talend.com/resource/big-data-transfer.html https://www.elastic.co/ https://hive.apache.org/ http://flume.apache.org/
Data visualisation	R Cytoscope Cicos Gephi IBMMany Eyes GraphViz Tableau PanXpan FusionCharts	Data visualisation	http://cran.r-project.org/ http://www.cytoscape.org/ http://circos.ca/ https://gephi.github.io/ http://www-01.ibm.com/software/analytics/many-eyes/ http://www.graphviz.or/ http://www.tableausoftware.com/ https://www.panxpan.com http://www.fusioncharts.com/

(source: https://www.researchgate.net/publication/309740042_Big_data_in_food_safety_An_overview)

EXAMPLES OF DATA ANALYSIS METHOD



Analysis method	Analysis method type	Applications
Recommendation system	Collaborative Filtering	Amazon.com (Linden et al., 2003a)
	Content-based filtering	Netflix (Koren, 2008); MovieLens (Miller et al., 2003)
	Heuristics	VERSIFI Technologies (Parikh and Zitnick, 2011)t
	Hybrid approaches	
Machine learning	Auto Encoder	Speech recognition (Liu and Yang, 2015); (Hu and Nie, 2016)
	Restricted Bolzmann Machine	Natural Language Processing (Agerri et al., 2015)
	Bayesian networks	Protein-protein interaction network (Chen and Qiao, 2015);
	Neural networks	Disease gene priorization (Li et al., 2012).
	Transfer Learning	Food fraud prediction (Bouzembrak and Marvin, 2016; Marvin et al., 2016)
	Manifold Learning	
	Topological analysis	
	Guilt-by-association	
	Shortest path analysis	

(source: https://www.researchgate.net/publication/309740042_Big_data_in_food_safety_An_overview)

Thanks!

CREDITS: This presentation template was created by **Slidesgo**,
including icons by **Flaticon**, and infographics & images by **Freepik**