Big Data & Public Health

Mr. Sam Ng

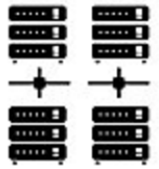
Automated Systems (H.K.) Limited

Agenda

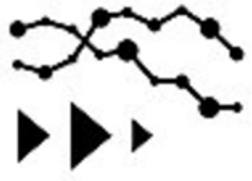
- Big Data Overview
- Data Analytics
- Data Preparation and Data Quality
- Data Masking

Big Data Overview

Capabilities



Big Data: extracting insight from large amounts of unstructured data



Fast Data: transforming, filtering, aggregating, and correlating streams of data in near real-time

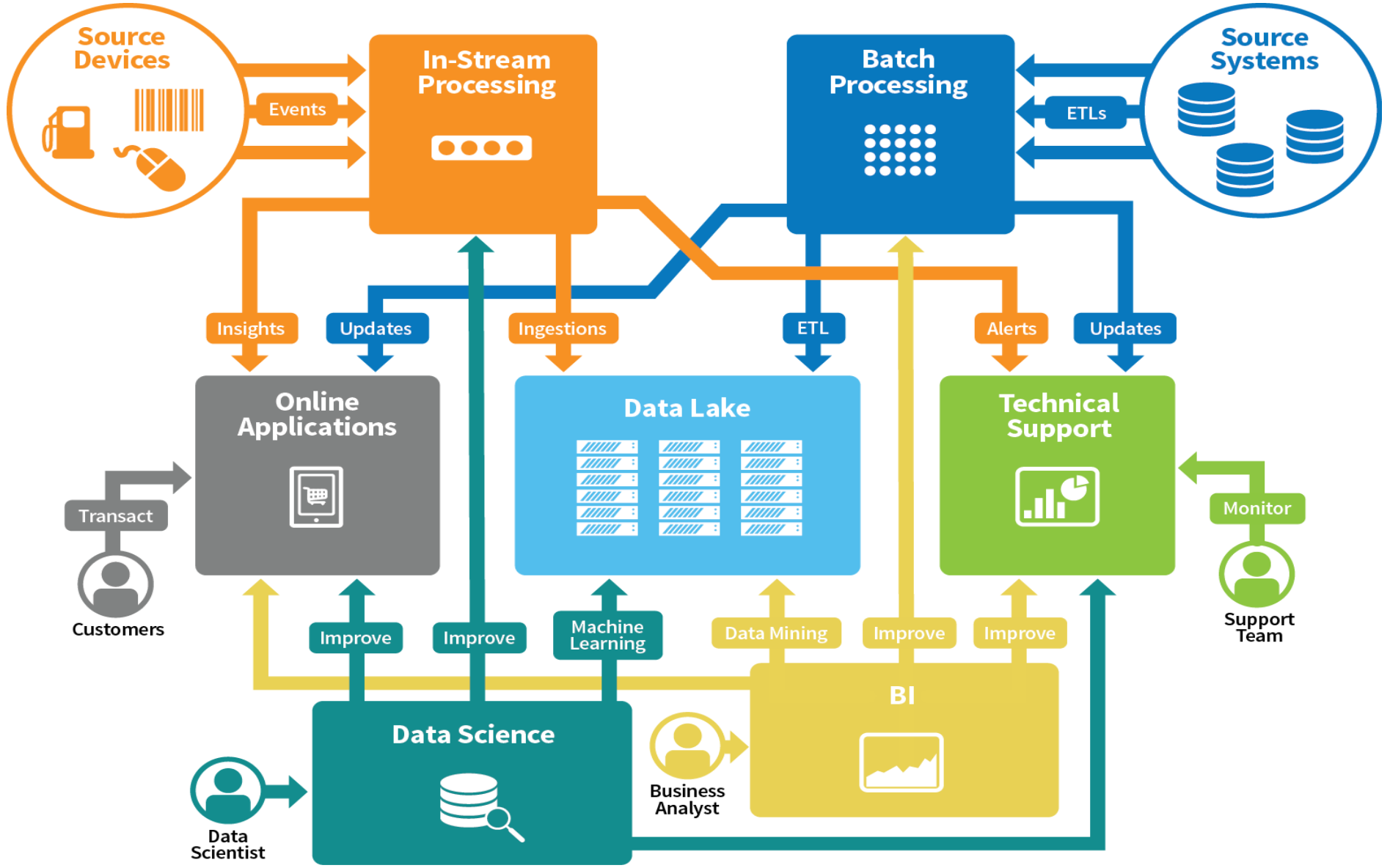
Skills

Visualization and Reporting	Data Science and Machine Learning	Data Engineering	Infrastructure Engineering

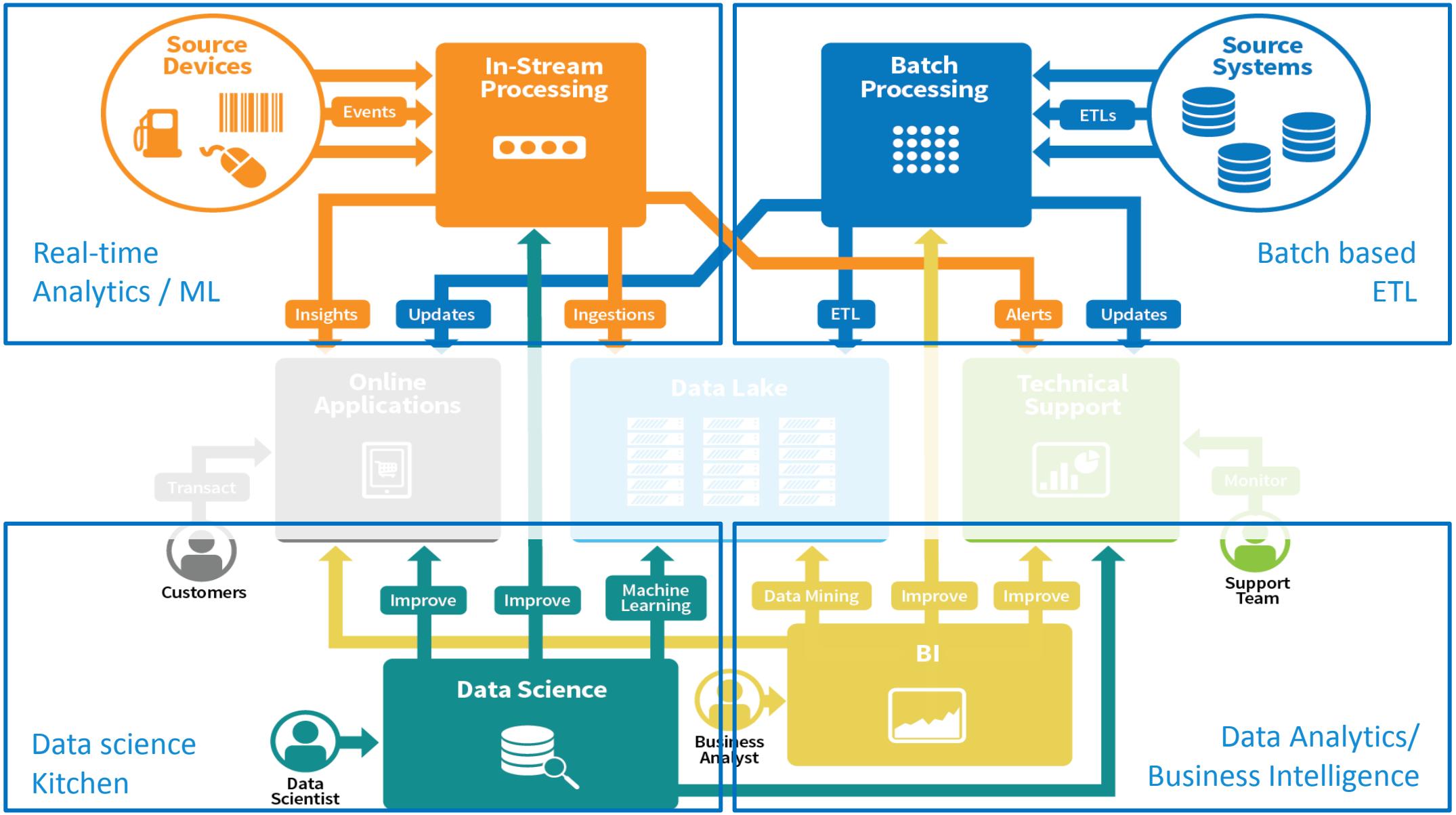
Technology Enablers

Internal Reporting	Interactive exploration	In-Memory Databases	CI/CD and DevOps
Customer-facing Charting	**Data Analytics**	Data Modeling	Performance Engineering
Rich Data Interfaces	Deep Learning	Data Lakes	Cloud Migration
	Integration with statistical tools	**ETL**	Managed Services
		Data Governance	Security

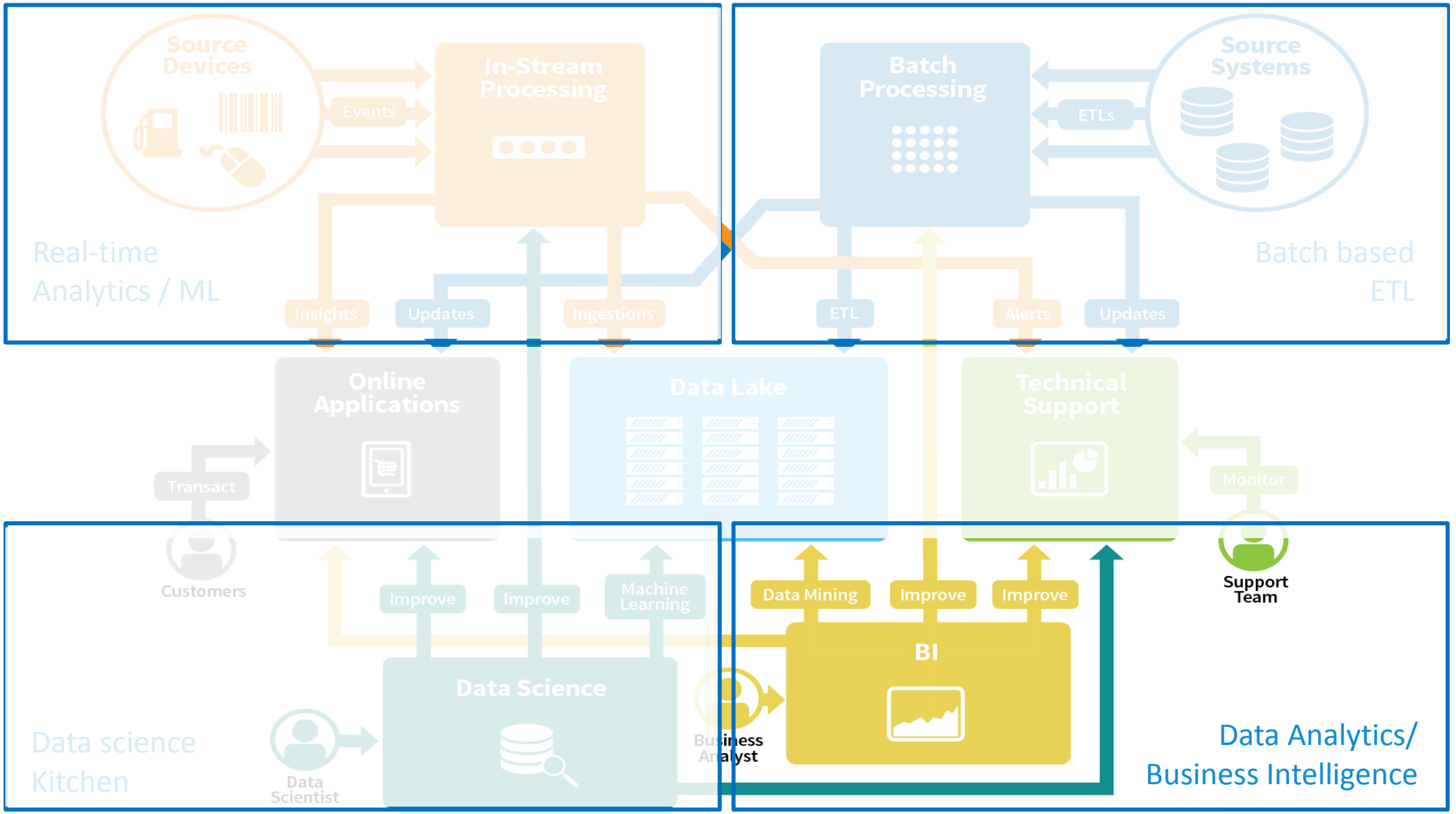
Full coverage of modern big data landscape



Full coverage of modern big data landscape



Data Analytics



Data Analytics

Descriptive Analytics

What has happened?

Uses data aggregation and data mining to provide insight into the past

Predictive Analytics

What could happen?

Uses statistical models and forecasts techniques to estimate the future

Prescriptive Analytics

What should we do?

Uses optimization and simulation algorithms to advice on possible outcomes

Descriptive Analytics

Descriptive Analytics → Insight into the past

Descriptive analysis or statistics “describe”, or summarize raw data and make it interpretable by humans. They are analytics that describe the past.

Examples: all kinds of data mining / analytical reports / dashboards

- Communicable & Non-communicable Diseases report
- In-patient Discharge per hospitals
- Population Demographics reports
- Birth Weight Statistics reports

Example: HealthyHK, Department of Health

HealthyHK, Department of Health
The Government of the Hong Kong Special Administrative Region

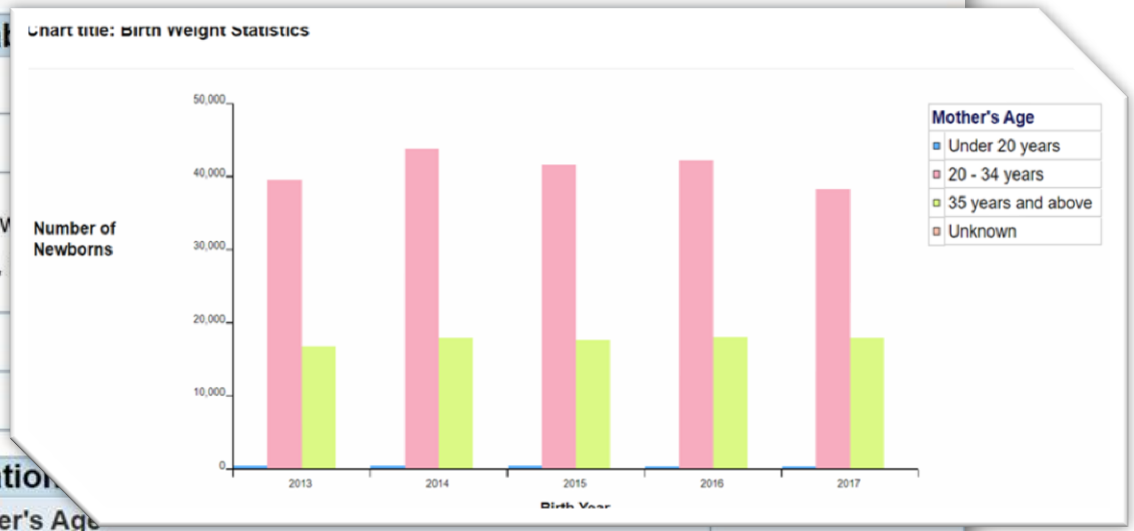
HONG KONG

Welcome to Instant Query!

Birth Weight Statistics

Birth Weight Statistics Report

Specified row and column variab	
Birth Year :	2013, 2014, 2015, 2016, 2017
Sex of Baby :	Male, Female, Other, Unknown
Mother's Residential District :	Central & Western, Eastern (HK), Southern (HK), Wan Chai, Kowloon City, Kw North, Sai Kung, Sha Tin, Tai Po, Tsuen Wan, Tuen Mun, Yuen Long, Marine,
Mother's Age :	Under 20 years, 20 - 34 years, 35 years and above, Unknown
Plurality :	Singleton, Twins, Triplets, Quadriplets or More, Unknown



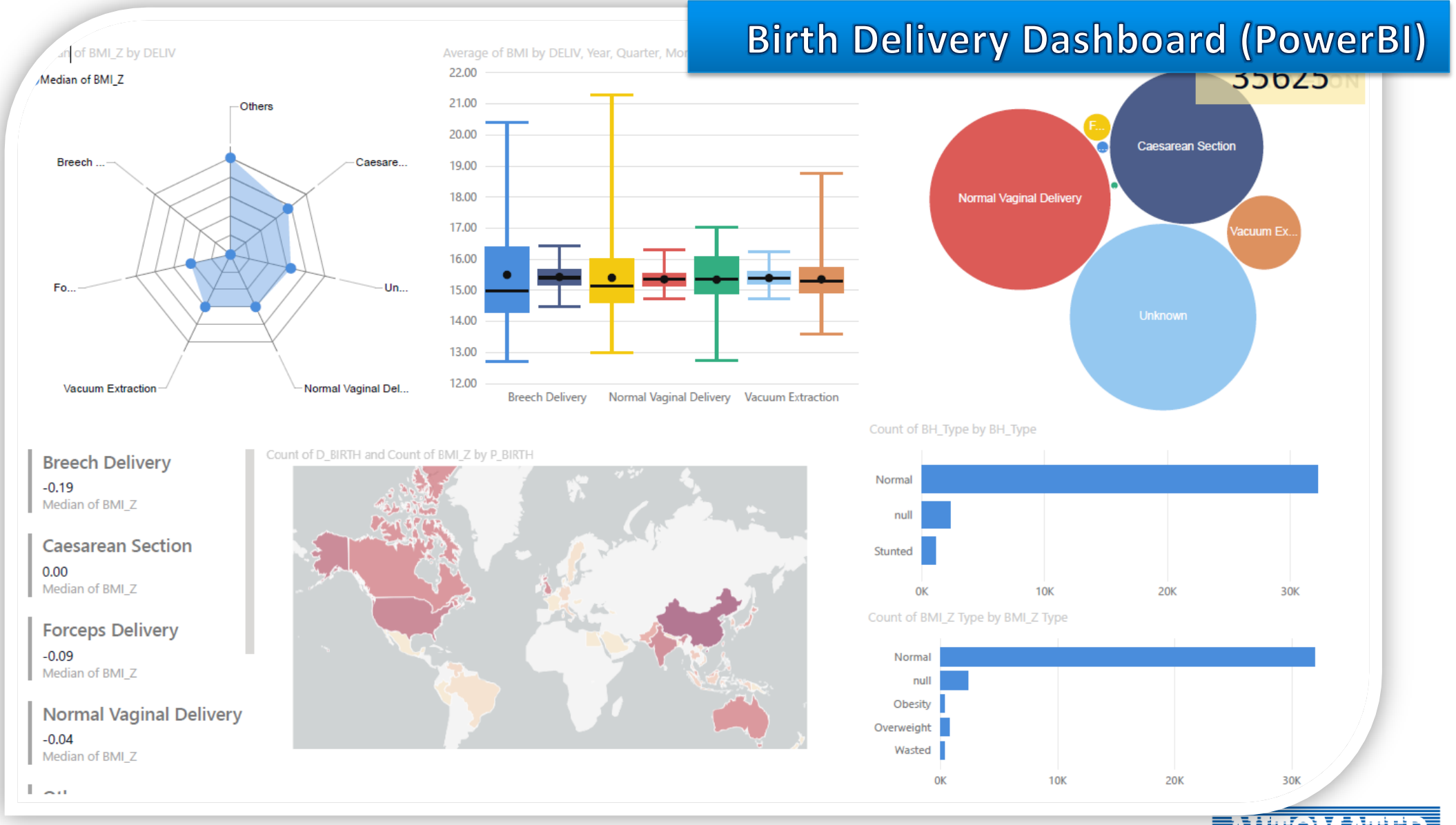
Results in Tabulation

Number of Newborns		Mother's Age				Row Total
		Under 20 years	20 - 34 years	35 years and above	Unknown	
Birth Year	2013	525	39,653	16,860	37	57,075
	2014	559	43,742	17,953	36	62,290
	2015	494	41,664	17,650	62	59,870
	2016	437	42,236	18,140	41	60,854
	2017	393	38,242	17,907	---	56,542
All Shown Birth Year		2,408	205,537	88,510	176	296,631

AUTOMATED ASL A member of the Teamsun Group

Example: PHIS, Department of Health

Birth Delivery Dashboard (PowerBI)



Predictive Analytics

Predictive Analytics → Estimate future

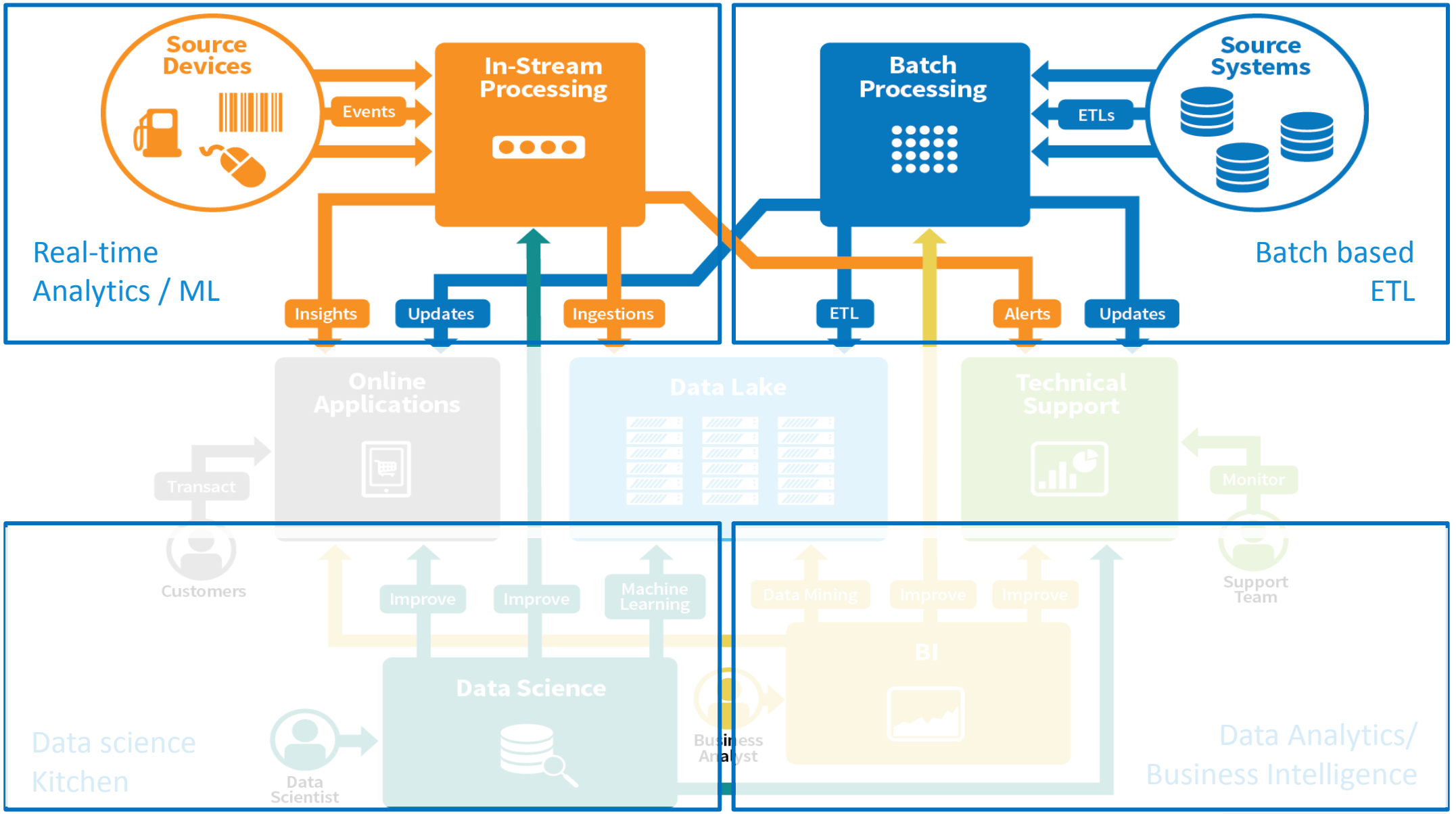
Predictive Analytics are about understanding the future and providing companies with actionable insights based on available data. They provide estimates about the likelihood of future outcomes.

Examples: all kinds of forecasting / machine learning / AI

- Population Projection
- Epidemic Outbreak Prediction

How to make Predictive Analytics success?

Data Preparation



 |

Data Preparation

Data Preparation → transform data into "ready for analysis"

Data preparation is the process of manipulating data into a form that is suitable for analysis. "Suitable" in this case means that the data is **clean, complete, and quantifiable**.

Analytics typically includes internal data sources, such as transactional systems, data warehouses and departmental data marts. Nowadays, analytics includes also external data sources, such as social media, environmental data (weather, air, water quality), WHO, etc.

Through **extract-transform-load (ETL)** tasks, it can transform the data sets into a form that's compatible with the needs of predictive analytics.

Data Preparation

Here are a few steps to prepare data for analytics:

1. **Streamline access to the data** — Building a robust data pipeline requires a way to quickly and efficiently refresh data sets used for predictive purpose. The data access part of the pipeline should be flexible enough to accept new formats and structures.

2. **Build an abundant data transformation toolbox** — Data size and complexity of the data transformation toolbox will grow over time. Common tasks such as sorting, merging, aggregating, reshaping and partitioning, etc.

3. **Statistical analysis** — During the data prep stage, it often needs to perform exploratory statistical analysis to gain deep familiarity with the data.

Data Quality

Data quality has become a major focus of public health programs in recent years, especially as the demand for accountability increases.

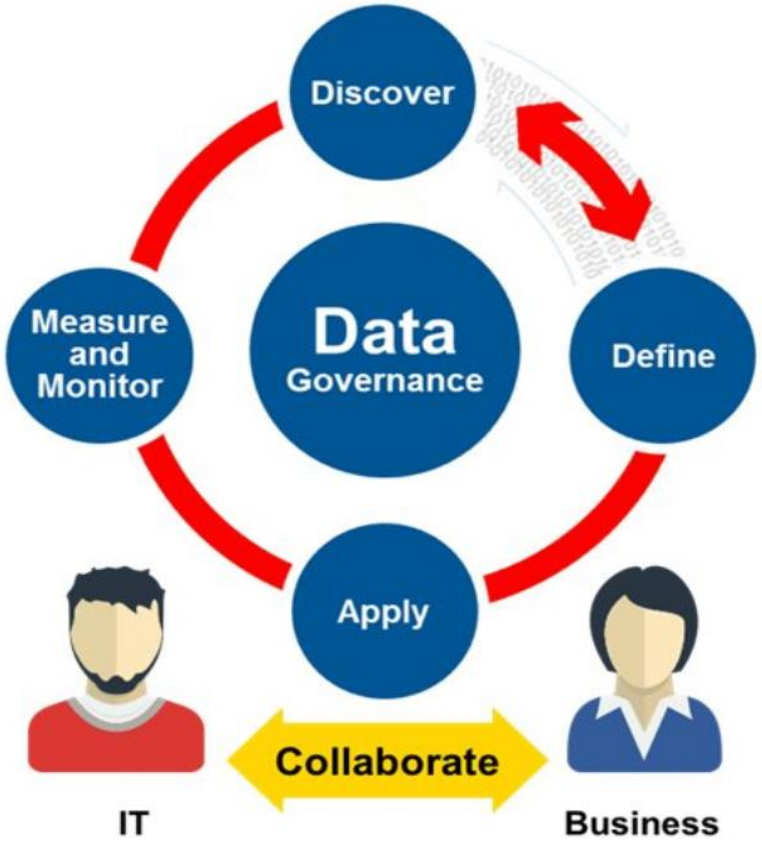
To fight against diseases such as AIDS, Tuberculosis, and Malaria must be predicated on strong Monitoring and Evaluation systems that produce quality data related to program implementation.

- increasingly seek tools to **standardize and streamline** the process of determining the quality of data;
- verify the quality of reported data, and assess the underlying data management and reporting systems for indicators.

An example is WHO and MEASURE Evaluation's Data Quality Review Tool

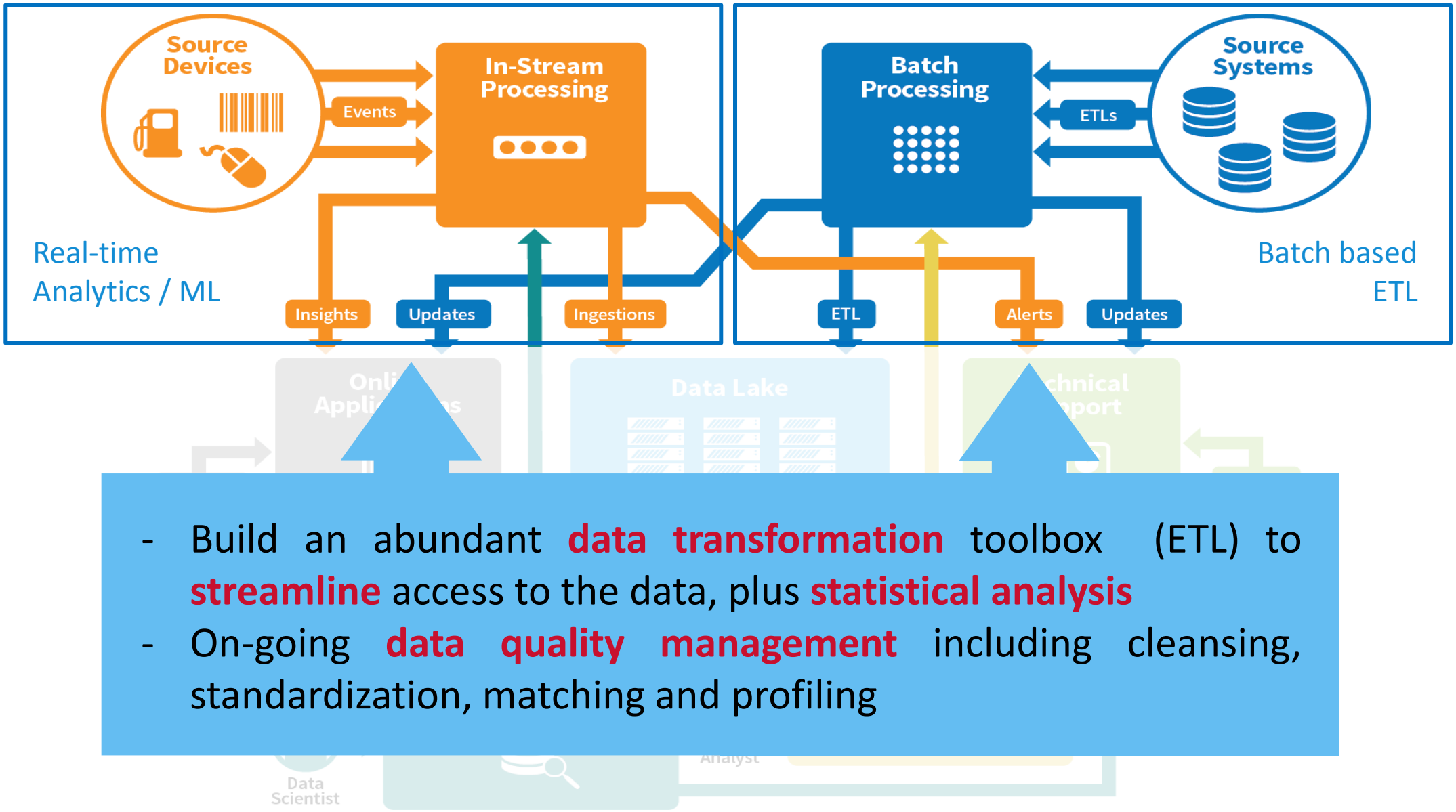
Data Quality

Examples of data quality solutions include **cleansing, standardization, matching, profiling**, and more. It helps to maintain and improve the quality of your data, ensuring that you can make predictive analytics success.



Data quality is NOT “one and done” process. Data quality management is **an ongoing cycle** that needs to be done on the front-end as data is coming, but also on the back-end on a regular basis to keep legacy data up to the highest quality, integrity, and consistency standards.

Data Preparation and Data Quality



- Build an abundant **data transformation** toolbox (ETL) to **streamline** access to the data, plus **statistical analysis**
- On-going **data quality management** including cleansing, standardization, matching and profiling

Data Masking

The extensive use of electronic health data has **increased privacy concerns**. While most healthcare organizations are conscientious in protecting their data in their databases, very few organizations take enough precautions to protect data that is shared with third party.

PII – Personally Identifiable Information – any data that could potentially identify a specific individual. Any information that can be used to distinguish one person from another and can be used for de-anonymizing anonymous data can be considered PII.

PHI – Protected Health Information – any information about health status, provision of health care, or payment for health care that can be lined to a specific individual.

Data Masking

Data masking – a process that scrambles data, either an entire database or a subset. Unlike encryption, masking is not reversible; And masked data is useful for limited purposes.

There are several types of data masking:

- **Persistent (static) data masking** masks data in advance of using it. Non production databases masked NOT in real-time.
- **Dynamic data masking** masks production data in real time
- **Data Redaction** – masks unstructured content (PDF, Word, Excel)

Data Masking

Masking can scramble individual data columns in different ways so that the masked data **looks like the original (retaining its format and data type) but it is no longer sensitive data**.

Traditionally, data masking has been viewed as a technique for solving a test data problem. However, the scope of data masking is extended to more broadly include **data de-identification in production, non-production, and analytic use cases**. The challenge is to do this while retaining business value in the information for consumption and use.

Conclusion: Data Governance

Discover

- Data Quality
- Rev
- Secure@ Source

Define

- Enterprise Data Integration
- Data Quality
- Master Data Management
- Dynamic Data Masking

Measure and Monitor

- Enterprise Data Integration
- Data Quality
- Master Data Management
- Dynamic Data Masking
- Proactive Monitoring
- Secure@ Source

Apply

- Enterprise Data Integration
- Data Quality
- Master Data Management
- Dynamic Data Masking

informatica

 |

AUTOMATED ASL A member of the Teamsun Group

Approach in Data Intelligence Realization

ASL will adopt an end-to-end service approach to help customers successfully realize their data platform. Starting with a detailed requirement study, our experienced consultants would lead the customer to think wider, think deeper and think further. Major services delivered by ASL in Big Data / Business Intelligence Area cover:

Makes Your Data Intelligent!

• Big Data Strategy & Architecture Framework	• Mobile BI, Cloud BI
• Reporting and Analytics	• Data Mining & Statistical Analysis
• Data Visualization	• Predictive Analytics
• Business Discovery	• Big Data Integration
• Enterprise Data Architecture	• Corporate Performance Management
• Data Warehouse & Modelling	• Budgeting, Planning, Consolidation
• Enterprise Data Governance	• Data & User Security
• Enterprise Data Integration / ETL	• Backup & Restore in BI / Big Data Platform
• Data Cleansing & Data Quality	• Maintenance Support

Thank you

Your **Trustworthy** and **Professional** IT Partner

