Data Cheat Sheet

Term	Concept	Description /Analogy	Illustration	Real world Example	Relevance to Ki
Systems of Record	An information storage system that is the authoritative data source for a given data element or piece of information	• Used as a data repository where the data object, as a whole or specific attributes of a data object, are maintained (data creation, updating, modifying, and deleting) • Helps organisations to manage big data by serving as a primary source	Click to view illustration	Electronic Health Records	
Systems of Insight	A system of insight facilitates organizing, transforming, and analysing your data through the consumption, collection, and analysis of data from the combined sources of traditional "systems of record" and "systems of engagement"	• By leveraging behaviour-driven insights, systems of insight apply advanced analytics to generate customer interaction data • Allows businesses to immediately deploy personalisation	Click to view illustration	Analyse business data and social feeds, such as social networks, customer service interactions, web clickstream data to predict potential customer behaviour and determine targeted ads	
Systems of Engagement	A system that manages and promotes user collaboration and interaction.	• A system of engagement overlays and complements an organisation's investment in a system of record by providing easy access to data, as well as easy to use applications that enable collaboration across your organisation	Click to view illustration	Enhance patient engagement through mobile notifications system	
Data Warehouse	A central repository of information that can be analysed to make more informed decisions	• Data flows into a data warehouse from transactional systems, relational databases, and other sources, typically on a regular cadence • Enables powerful analytics on huge volumes (petabytes and petabytes) of historical data in ways that a standard database cannot	Click to view illustration	Google BigQuery SQL Data Warehouse	
Data Lake	A storage repository that holds a vast amount of raw data in its native format until it is needed	• Enables storage of all structured or unstructured data at scale • Uses a flat architecture to store data • Ability to run different types of analytics to guide better decisions	Click to view illustration	Azure Data Lake	

Term	Concept	Description /Analogy	Illustration	Real world Example	Relevance to Ki
Data Mesh	Data mesh is an architectural paradigm that embraces the ubiquity of data in the enterprise by leveraging a domain-oriented, self-serve design	• Used to stitch together data held across multiple data silos • Allows for a federated data architecture by bringing autonomy to data domains, while providing standardisation through central capabilities	Click to view illustration	JPMorgan Chase build a data mesh architecture to eable data sharing across the enterprise while giving data owners the control and visibility they need to manage their data effectively	
Centralised data platform approach	Data brought into a central platform to enable insights and analytics to allow users to use the data for strategic purposes	• Enables a cohesive view of data from multiple sources • Facilities better optimisation through better tracking, improved data integrity, reduced data redundancy, and more	Click to view illustration	By taking a centralised data platform approach, the Bank of England was able to break down data siloes, improve speed of reporting by more than 500x from days to minutes, and ensure data is reliable and compliant	
Data streaming	The process of transmitting, ingesting, and processing data continuously rather than in batches	• Enables generation of analytic results in real time as streamed data can be processed and analysed as it arrived • Benefits include easy data scalability, detection of patterns in time-series, and more	Click to view illustration	Real-time stock trades, up-to-the-minute retail inventory management, social media feeds, multiplayer game interactions, and ride-sharing apps	
Data model	The process of creating a visual representation of either a whole information system or parts of it to communicate connections between data points and structures	• Illustrates the types of data used and stored within the system, the relationships among these data types, the ways the data can be grouped and organised and its formats and attributes	Click to view illustration	*See below data modelling techniques for examples*	
Hierarchical Model	A tree-like structure where there is one root node (parent node) and the other child nodes are sorted in a particular order	• Simple and rigid structure • Very rarely used nowadays as retrieving and accessing data is difficult in a hierarchical database	Click to view illustration	Real-world relationships e.g. food recipes, sitemap of a website	
Object-oriented Model	A data model that consists of a collection of objects, each with its own features and methods	• Communicates while supporting data abstraction, inheritance, and encapsulation	Click to view illustration	db4o, Smalltalk and Cache	

Term	Concept	Description /Analogy	Illustration	Real world Example	Relevance to Ki
Network Model	A data model in which each record can be linked with multiple parent records	• Similar to hierarchical model but easier to convey complex relationships • Useful for projects of a complex nature or where activities are subject to a considerable degree of uncertainty in performance time	Click to view illustration	Modelling protein-protein interaction during the drug discovery process for the pharmaceutical industry	
Entity-relationship Model	A high-level data model that defines data entities (e.g. concept, data, object) and their relationship for a specified software system	• Represented in a diagram called an entity-relationship diagram, consisting of Entities, Attributes, and Relationships • Provides a better view of the data that is easy for stakeholders and developers to understand	Click to view illustration	Database of a school: Entities – 'Teacher', 'Department' Attributes – (of 'Teacher') Salary, Age, ID Relationship – 'Teacher' works for 'Department'	
Relational Model	A data model used to describe the different relationships between the entities	• Most widely used data model • Minimal complexity and provides a clear overview of the data	Click to view illustration	MySQL, PostgreSQL – used to run queries and create reports e.g. consolidated customer statements	
Data Pipelines	A set of data processing elements connected in series, where the output of one element is the input of the next one	• Enables the flow of data from an application to a data warehouse, from a data lake to an analytics warehouse, or into a system	Click to view illustration	Data pipelines are used to perform predictive analysis e.g. a production department can use predictive analytics to know when the raw material is likely to run out, and it could also help forecast which supplier could cause delays	

ACORD Reference Architecture	A series of seven interrelated industry models, or facets, which use different views to define the nature of the insurance industry	• Consists of:
		  • Business glossary
		  • Informational models
		  • Data models
		  • Capability models
		  • Component models
		  • Process models
		  • Products frameworks
		• Implementation of ACORD standards improves data quality and improves data quality and flow, increase efficiency, and realize billion-dollar savings
			Click to view illustration	Official website	

Term	Concept	Description /Analogy	Illustration	Real world Example	Relevance to Ki
Batch-based data pipeline	Data pipelines which are executed manually or recurringly	• In each run all data is extracted from the data source, operations are applied to the data and the processed data is published to a data sink • Execution time ranges from a few minutes to a few hours depending on the quantity of data	Click to view illustration	Payroll, billing, low frequency reports based on historical data	
Streaming data pipeline	Data pipelines which are executed continuously all the time	• Consumes streams of messages, apply operations to the messages (such as transformations and filters) and publishes the processes messages to another stream • Enables real-time processing	Click to view illustration	Netflix uses streaming data pipelines to cater for its requirements of real time evet-based processing and extensive support for customisation of windowing	
Lambda architecture	A data-processing architecture of three layers: batch processing, speed processing and a serving layer for responding to ad hoc queries	• Designed to handle massive quantities of data by taking advantage of both batch and stream processing methods • Benefits include tolerance to human errors and hardware crashes, and scalability and quick response time	Click to view illustration	Yahoo uses lambda architecture for running analytics on its advertising data warehouse	
Event-driven architecture	A software architecture paradigm promoting the production, detection, consumption of, and reaction to events (an event is a change in state or an update e.g. an item being placed in a shopping cart)	• Uses events to trigger and communicate between decoupled services • Has three key components: event producers, event routers and event consumers- producers publish an event to the router, which filters and pushes the event to consumers • Common within modern applications built with microservices	Click to view illustration	If a customer returns an item, this is logged by the stock management system (event producer), the event router pushes this to the finance system (event consumer) which updates to reflect this	

Term	Concept	Description /Analogy	Illustration	Real world Example	Relevance to Ki
Structured data	Data that has been predefined and formatted to a set structure before being placed in data storage, which is often referred to as schema-on-write	• Easily used by business users/ ML algorithms to draw business outcomes • Quick retrieval and less storage space required than unstructured data	Click to view illustration	Credit card numbers, dates, address	
Unstructured data	Data stored in its native format and not processed until it is used	• A variety of file formats that cannot be displayed in rows/columns as a relational database • Requires data science expertise/specialised tools (e.g. NLP) to extract value	Click to view illustration	Emails, photos, social media posts, text, audio files	