

Data Activities

Upstream:
Data Production

The Promise of Data. Many experts believe data (and particularly big data) hold the key to the future because of their ability to reveal patterns and connections that significantly improve lives from secure self-driving cars to more effective pharmaceutical treatments to more reliable weather forecasts enabling farmers to get better yields or predicting drought conditions. To understand how to harness the benefits of data, the starting point is to understand what data are, who generates data, and who collects data.

Data Creation. Data volumes have skyrocketed. From 2010 to 2020, the amount of data created, captured, copied, and consumed in the world increased from 1.2 trillion gigabytes to 59 trillion gigabytes, an almost 5,000% rate of growth! IBM estimates there are 2.9 million emails sent every second, 375 megabytes of data consumed by households daily, 20 hours of video uploaded to YouTube every minute, 24 petabytes of data processed per day by Google, and 73 products ordered on Amazon per second. More data was generated in the last two years than in the entire human history before that. The total amount of data created, captured, copied, and consumed globally is forecast to increase rapidly, reaching 180 zettabytes in 2025. We are swimming in a world of data.

Data Creators. Every individual, business enterprise, and government agency everywhere generates data. Individuals constantly generate data, primarily of a personal nature. On Google alone, people submit 40,000 search queries per second, which amounts to 1.2 trillion searches yearly! Each minute, 300 new hours of video show up on YouTube. That's why there are more than 1 billion gigabytes (1 exabyte) of data on Google's servers! People share more than 100 terabytes of data on Facebook daily. Every minute, users send 31 million messages and view 2.7 million videos. Smart devices (for example, fitness trackers, sensors, and Amazon Echo) produce 5 quintillion bytes of data daily.

Every business generates data (a) through its internal support functions (e.g., human resources, procurement, legal, accounting, R&D, sales and marketing) that tends to be similar across all business sectors and (b) arising from operations that are unique to its business sector (i.e., the products and services the company sells), such as healthcare (health insights, data on the effectiveness of different drug treatments, and improvements in emergency room care), banking (customer account balances, and loan delinquencies), entertainment media (the TV shows subscribers watched during peak viewing hours), retail (customer profiles and purchase histories and habits), energy and utility industries (sensors that indicate turbine and engine performance), construction (building construction sequencing, and subcontractor scheduling), and transportation (train conditions and fuel consumption).

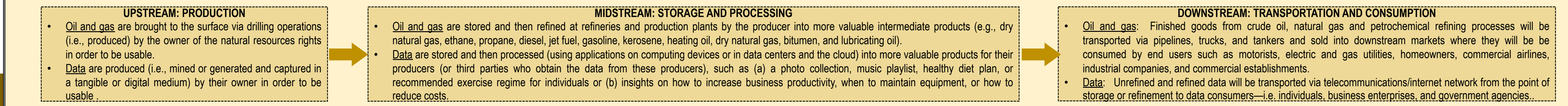
In the U.S., the federal government is perhaps the most prolific generator of data, including weather, employment, and economic statistics, surveillance footage of foreign troop movements, the flight paths of asteroids and comets, the amount of government student loans outstanding, and data on the incidence of disease.

Data Collection. Although individuals, businesses and government agencies generate data for themselves (original data generation), each data generator is involved in collecting data from the other data generators, which itself is a form of data generation (secondary data generation). For example, a business will collect personal data from its customers in order to establish an online banking account, the government will request data from a pharmaceutical company to determine whether to grant approval for a new drug, and individuals will collect data from the government or a business in order to initiate litigation.

Data as a
Natural
Resource

If the subjects of data generation, storage and processing, computing devices, information technology infrastructure for data centers and the cloud, and telecommunications networks are confusing to you, it helps to think of data in similar terms to the oil and gas industry because data are as much a natural resource as crude oil or natural gas found in Deepwater Offshore Gulf of Mexico or coal mined beneath the earth's surface in Wyoming or West Virginia.

There are **upstream activities** (exploration, mining, and production), **midstream activities** (storage, refining and processing), and **downstream activities** ("last mile pipeline distribution," retail sales, and consumption of finished goods).



GOVERNMENT DATA GENERATION AND COLLECTION

BUSINESS DATA GENERATION AND COLLECTION

INDIVIDUAL DATA GENERATION AND COLLECTION

Midstream:
Data Storage & Processing

Storage. Once generated, data will be stored until needed (a) on disks embedded in stationary and mobile computing devices (e.g., desktop and laptop computers, tablets, smartphones, digital assistants, wearables/watches, and fitness trackers), (b) in data centers operated by governments and business enterprises, and (c) in the cloud (i.e., mega-data centers operated by cloud service providers). Surprisingly, only a small percentage of newly created data (2%) is kept. Nonetheless, in line with the rapid growth of the data volume, the installed base of storage capacity is forecast to increase at a compound annual growth rate of 19.2%.

Computing (i.e., Analyzing and Processing). Over 99% of collected data never gets used or analyzed. Despite this tremendous waste of data, data that are ultimately used will be processed into more valuable products for their owners, such as (a) for individuals, a photo collection or recommended music playlist, healthy diet plan, or exercise regime for individuals, or (b) for businesses, insights on how to increase productivity and reduce costs, when to repair equipment, what goods to produce (and the price to sell them), or whether fraud may be occurring.

Applications. Big data is only as useful as the ability to read it. Therefore, data generators and collectors need tools to analyze and read the data. Businesses use tools to extract data from business systems and integrate it into a repository, such as a data warehouse. Once in the warehouse, the data can be analyzed. Analytical tools range from spreadsheets with statistical functions to enterprise resource planning systems (ERP), customer relations management programs (CRM), payroll tools, and operational systems.

Edge to Core to Cloud. The applications can be located (a) on-device (i.e., on the same computing device where the data are stored), (b) on-premises (i.e., in a data center maintained by an enterprise at a general or core location), (c) at the edge (i.e., at the location near where the data are generated), (d) in the cloud (i.e., a data center operated by a cloud service provider, where over 30% of all stored data is uploaded), or (e) any combination of the above. Because the data landscape is more dispersed than ever, the modern organization requires IT solutions that capture and analyze data as they move from "edge to core to cloud."

ENTERPRISE COMPUTING AND STORAGE

...IN THE CLOUD

...ON-PREMISES

...AT THE EDGE

PERSONAL COMPUTING AND STORAGE

...ON DEVICE

Downstream:
Data Transport & Consumption

Round-Trip Process. Raw, unprocessed data will be transported from an individual's device (whether acting alone or for a business or government entity) to a data center, private cloud or public cloud and back again as refined data. This cycle is essentially a "round-trip" process, where data is effectively mined, shipped, refined, and shipped again. Although the round-trip process typically occurs in the blink of an eye, the transport of data (as with any shipping process that involves logistics) takes time.

The Internet. Data (called messages) will be transported by its sender to a recipient along the Internet, a worldwide computer network (owned and operated by Internet service providers) that transmits a variety of data and media across interconnected devices. In the year 2000, only 52% of U.S. adults used the Internet. That number jumped to 90% in 2020. Approximately 7.5 billion people are projected to use the Internet by 2030 when over 500 billion devices will be connected to the Internet.

DATA TRANSMISSION

DATA CONSUMPTION


INDIVIDUAL DATA GENERATORS AND COLLECTORS

RESTAURANT

PERSONAL COMPUTING DEVICE MAKERS

DATA CONSUMPTION

DATA CONSUMPTION



The icon depicts three distinct elements: on the left, a silhouette of a person wearing a peaked cap and a uniform, likely representing a government official or law enforcement; in the center, a silhouette of an industrial factory with multiple windows and a tall smokestack emitting a plume of smoke; on the right, a silhouette of a family consisting of two adults and two children standing under a simple roofline.

Data Activities

Upstream: Data Production

The Promise of Data. Many experts believe data (and particularly big data) hold the key to the future because of their ability to reveal patterns and connections that significantly improve lives from secure self-driving cars to more effective pharmaceutical treatments to more reliable weather forecasts enabling farmers to get better yields or predicting drought conditions. To understand how to harness the benefits of data, the starting point is to understand what data are, who generates data, and who collects data.

Data Creation: Data volumes have skyrocketed. From 2010 to 2020, the amount of data created, captured, copied, and consumed in the world increased from 1.2 trillion gigabytes to 59 trillion gigabytes, an almost 5,000% rate of growth! IBM estimates there are 2.9 million emails sent every second, 375 megabytes of data consumed by households daily, 20 hours of video uploaded to YouTube every minute, 24 petabytes of data processed per day by Google, and 73 products ordered on Amazon per second. More data was generated in the last two years than in the entire human history before that. The total amount of data created, captured, copied, and consumed globally is forecast to increase rapidly, reaching 180 zettabytes in 2025. We are swimming in a world of data.

Data Creators: Every individual, business enterprise, and government agency anywhere generates data. **Individuals** constantly generate data, primarily of a personal nature. On Google alone, people submit 40,000 search queries per second, which amounts to 1.2 billion searches yearly! Each minute, 300 new hours of video show up on YouTube. That's why there are more than 1 billion gigabytes (1 exabyte) of data on Google's servers! People share more than 100 terabytes of data on Facebook daily. Every minute, users send 31 million messages and view 2.7 million videos. Smart devices (for example, fitness trackers, sensors, and Amazon Echo) produce 5 quintillion bytes of data daily.

Every **business** generates data (a) through its internal support functions (e.g., human resources, procurement, legal, accounting, R&D, sales and marketing) that tends to be similar across all business sectors and (b) arising from operations that are unique to its business sector (i.e., the products and services the company sells), such as healthcare (health insights, data on the effectiveness of different drug treatments, and improvements in emergency room care), banking (customer account balances, and loan delinquencies), entertainment media (the TV shows subscribers watched during peak viewing hours), retail (customer profiles and purchase histories and habits), energy and utility industries (sensors that indicate turbine and engine performance), construction (tracking construction sequencing, and subcontractor scheduling), and transportation (train conditions and fuel consumption).

In the U.S., the federal **government** is perhaps the most prolific generator of data, including weather, employment, and economic statistics, surveillance footage of foreign troop movements, the flight paths of asteroids and comets, the amount of government student loans outstanding, and data on the incidence of disease.

Data Collection: Although individuals, businesses and government agencies generate data for themselves (*original data generation*), each data generator is involved in collecting data from the other data generators, which itself is a form of data generation (*secondary data generation*). For example, a business will collect personal data from its customers in order to establish an online banking account, the government will request data from a pharmaceutical company to determine whether to grant approval for a new drug, and individuals will collect data from the government or a business in order to initiate litigation.

Midstream:

Storage & Processing

Storage. Once generated, data will be stored until needed (a) on disks embedded in stationary and mobile computing devices (e.g., desktop and laptop computers, tablet smartphones, digital assistants, wearables/watches, and fitness trackers), (b) in data centers operated by government agencies and business enterprises, and (c) in the cloud (i.e., mega-data centers operated by cloud service providers). Surprisingly, only a small percentage of newly created data (2%) is kept. Nonetheless, in line with the rapid growth of the data volume, the installed base of storage capacity is forecast to increase at a compound annual growth rate of 19.2%.

Computing (i.e., Analyzing and Processing). Over 99% of collected data never gets used or analyzed. Despite this tremendous waste of data, data that are ultimately used will be processed into more valuable products for their owners, such as (a) for individuals: photo collection or recommended music playlist, healthy diet plan, or exercise regime; (b) for businesses, insights on how to increase productivity and reduce costs, when to repair equipment, what goods to produce (and the price to sell them) whether fraud may be occurring.

Applications. Big data is only as useful as the ability to read it. Therefore, data generators and collectors need tools to analyze and read the data. Businesses use tools to extract data from business systems and integrate it into a repository, such as a data warehouse. Once in the warehouse, the data can be analyzed. Analytical tools range from spreadsheets with statistical functions to enterprise resource planning systems (ERP), customer relations management programs (CRM), payroll tools, and operations systems.

Edge to Core to Cloud: The applications can be located (a) on-device (i.e., on the same computing device where the data are stored), (b) on-premises (i.e., in a data center maintained by an enterprise at a central or core location), (c) at the edge (i.e., at a location near where the data are generated), (d) in the cloud (i.e., a data center operated by a cloud service provider, where over 30% of all stored data is uploaded), or (e) a combination of the above. Because the data landscape is more dispersed than ever, a modern organization requires IT solutions that capture and analyze data as they move from "edge to core to cloud."

I Downstream:

Transport & Consumption

Round-Trip Process. Raw, unprocessed data will be transported from an individual source (whether acting alone or for a business or government entity) to a data center, private cloud or public cloud and back again as refined data. This cycle is essentially a "round-trip" process, where data is effectively mined, shipped, refined, and shipped again. Although the round-trip process typically occurs in the blink of an eye, the transportation of data (as with any shipping process that involves logistics) takes time.

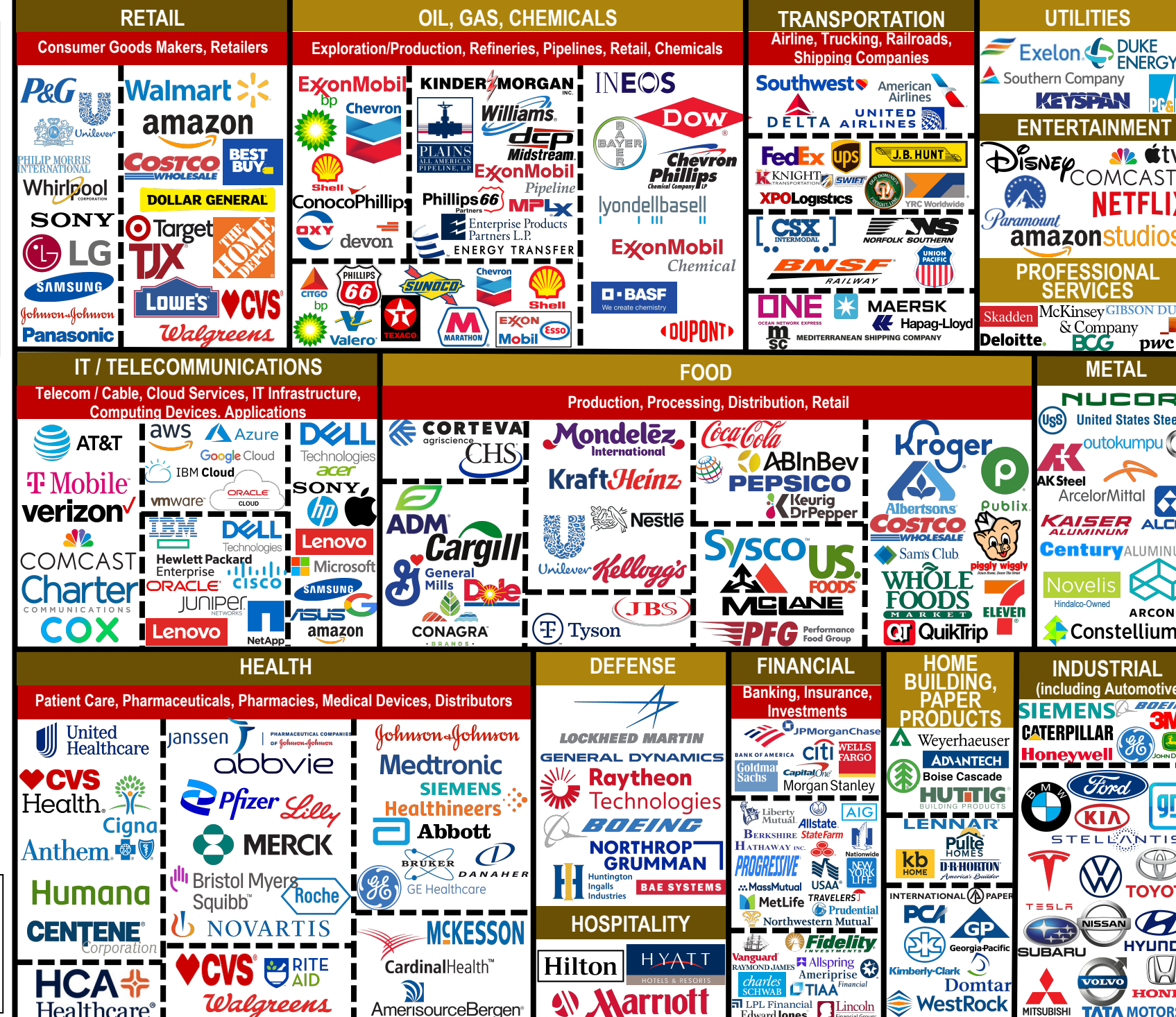
The Internet. Data (called messages) will be transported by its sender to a recipient along the internet, a worldwide computer network (owned and operated by internet service providers) that transmits a variety of data and media across interconnected devices. In the year 2000, only 52% of US adults used the internet. That number jumped to 90% in 2020. Approximately 7.5 billion people are projected to use the internet by 2030 when 500 billion devices will be connected to the internet.

Receipt and Consumption. Once refined data are received, government agencies, enterprises and individuals can delete the data, consume it (and then delete or store data), or work on the data (which restarts the data cycle).

GOVERNMENT DATA GENERATORS AND COLLECTORS



BUSINESS DATA GENERATORS AND COLLECTORS

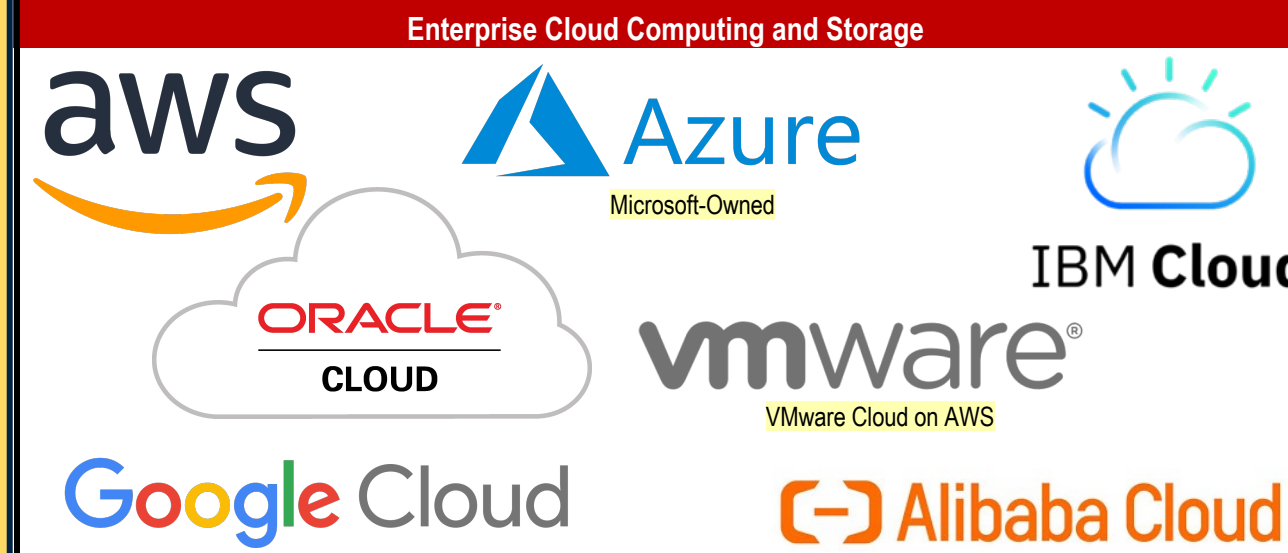


INDIVIDUAL DATA GENERATORS AND COLLECTORS



ENTERPRISE COMPUTING INFRASTRUCTURE AND SOFTWARE COMPANIES

...IN THE CLOUD



Personal Cloud Computing and Storage



...ON-PREMISES



Plant Automation



AT THE EDGE



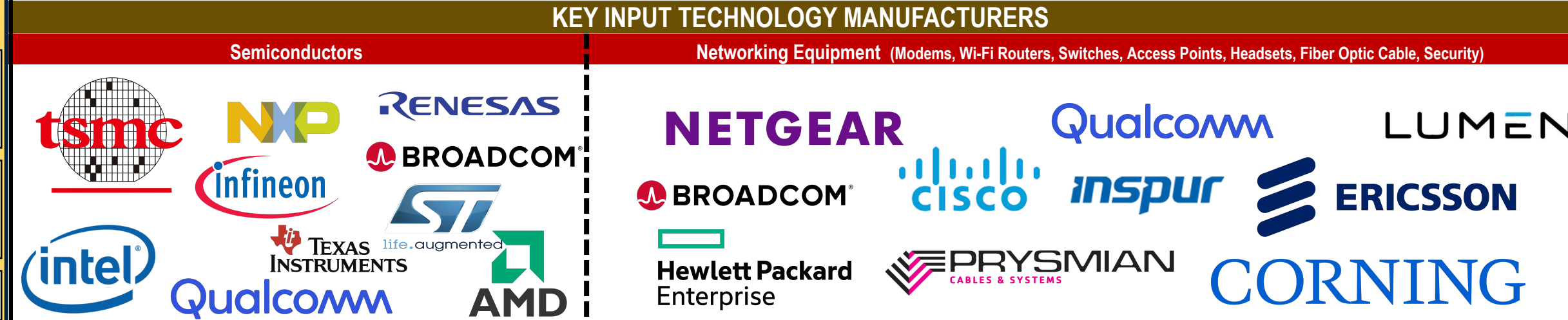
PERSONAL COMPUTING DEVICE MAKERS



Selected Developers of Business, Cloud-Hosted and Data Center Applications and Key Input Producers for IT Infrastructure



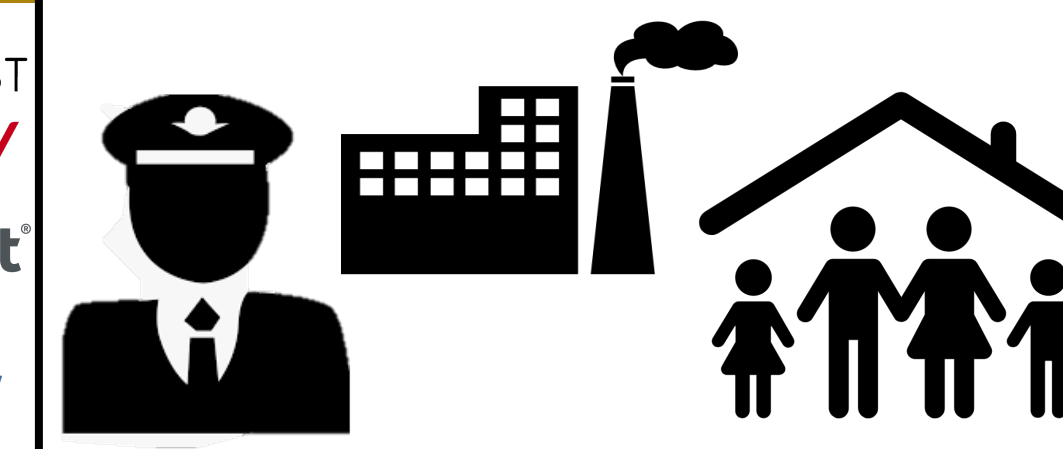
DATA TRANSMISSION COMPANIES



BROADBAND / INTERNET SERVICE PROVIDERS



DATA CONSUMPTION



Assumption. Once refined data are received, group members can delete the data, consume it (and then delete it), or store it (and then delete it). The data (which restarts the data cycle).

Interne

