FUJITSU

# Big Data – Eine Annäherung

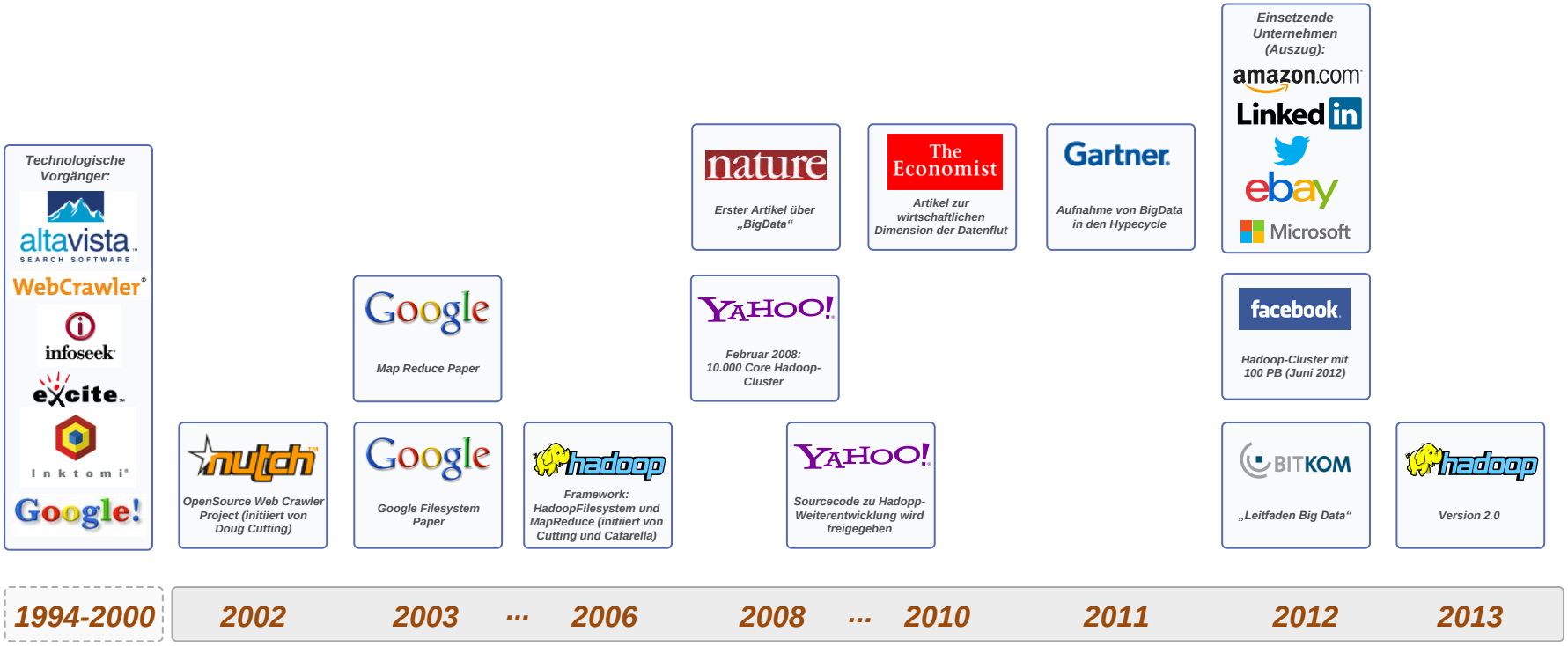
**Karsten Jansen – Fujitsu**

# Inhaltliche Schwerpunkte

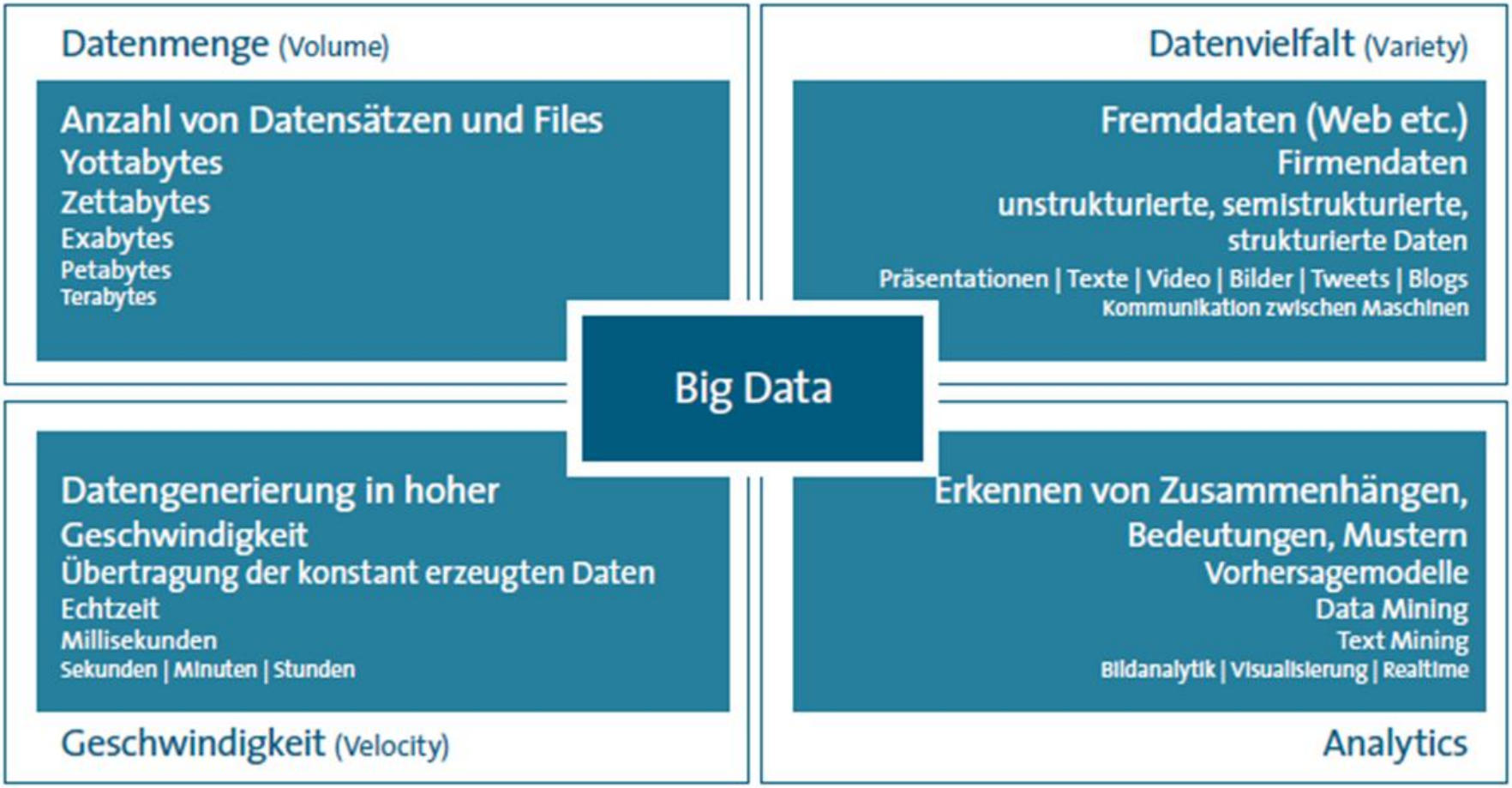
1. Wie alles begann –
   Eine technologische Einordnung
2. Fluch oder Segen –
   Auch Big Data hat ein Janus-Gesicht
3. Öffentliche Verwaltung in Deutschland –
   Realistische Ansätze in Nutzung und Umgang

# Morgens um zehn in Deutschland …

# Wie alles begann …



# Ein paar Grundlagen und Begriffe …



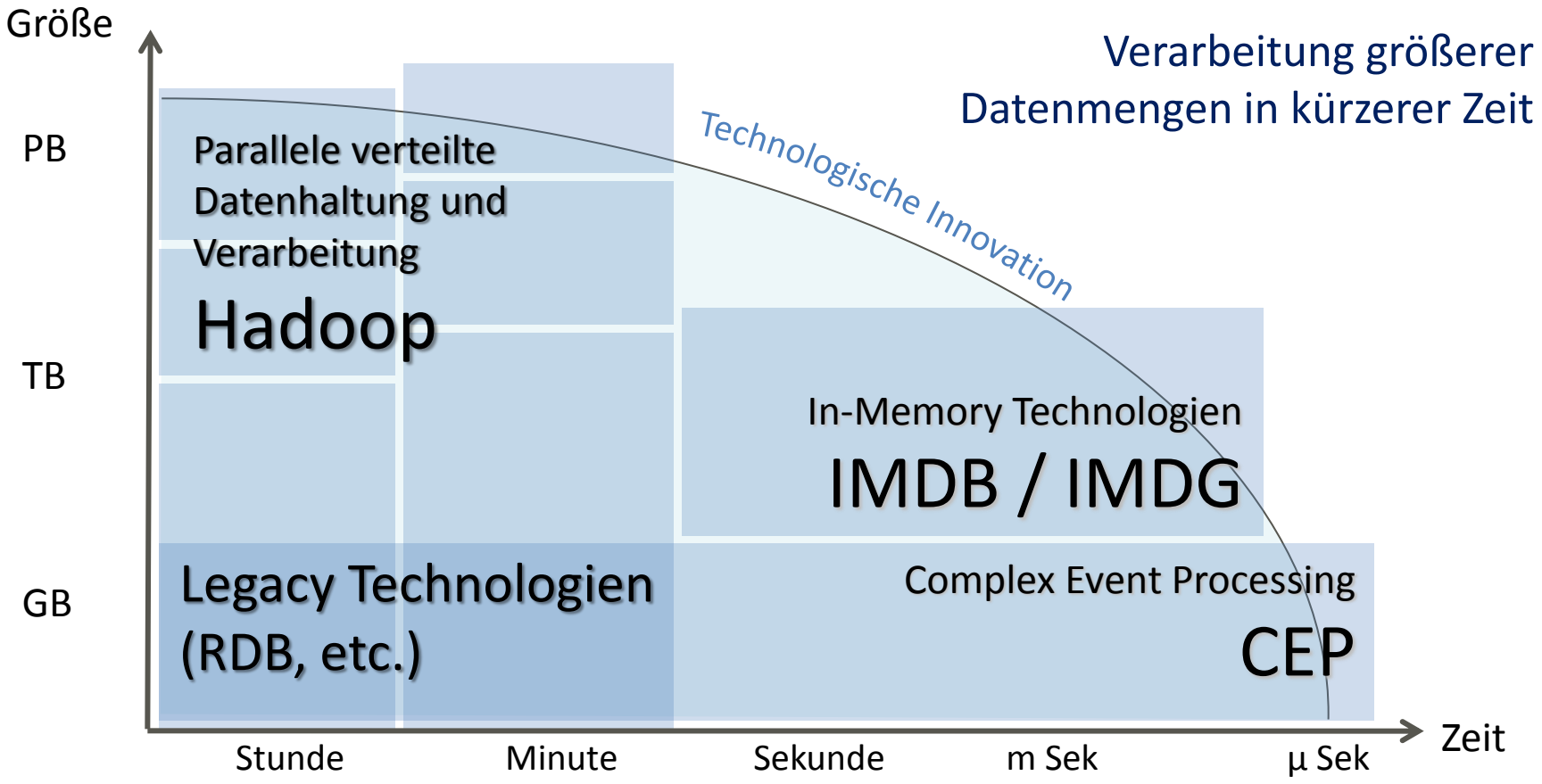

Quelle: BITKOM [2]

# Merkmale von BigData

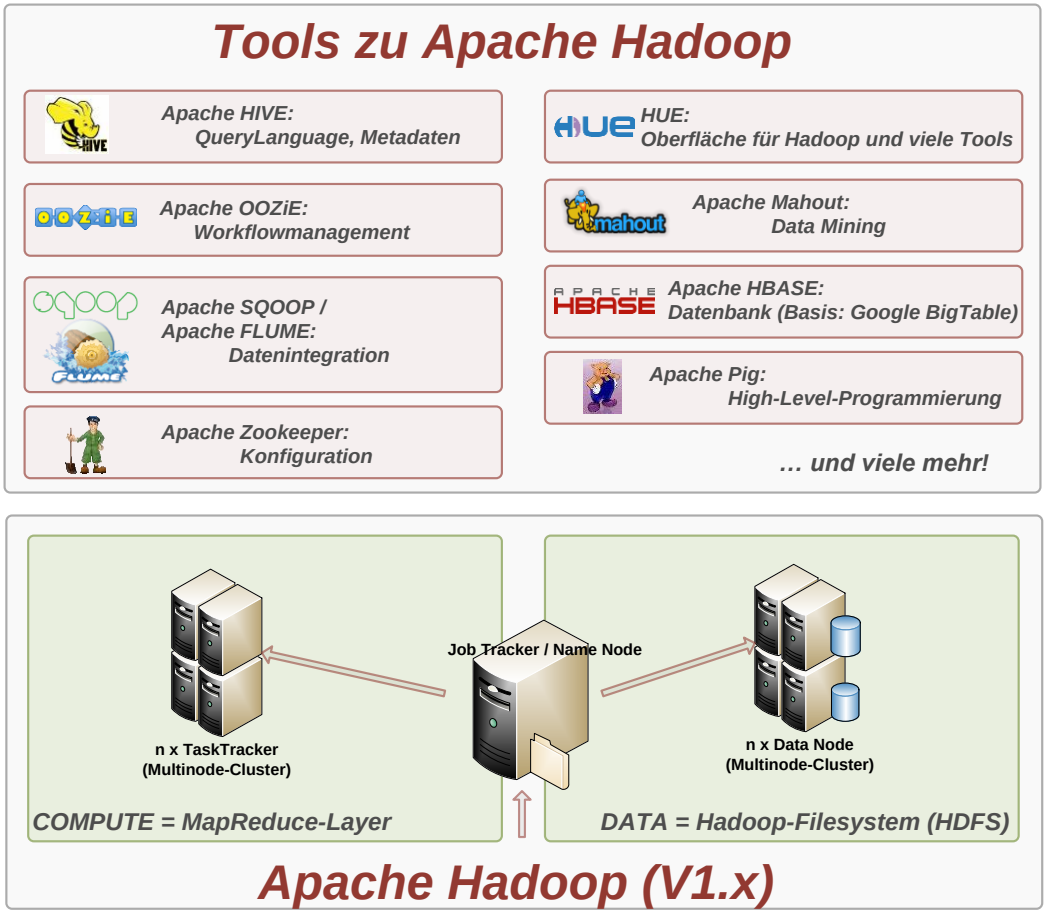
FUJITSU

- „Big Data bezeichnet die Analyse **großer Datenmengen** aus **vielfältigen Quellen** in hoher Geschwindigkeit mit dem Ziel, wirtschaftlichen Nutzen zu erzeugen“ (BITKOM).
- **Industriegetrieben:** Daten werden zu einem Wirtschaftsgut!
- Big Data stellt **Konzepte, Methoden, Technologien, IT-Architekturen** sowie **Tools** zur Verfügung.
- Big Data setzt da ein, wo **konventionelle Ansätze** der Informationsverarbeitung an **Grenzen** stoßen, die Flut zeitkritischer Informationen für die Entscheidungsvorbereitung zu bewältigen.
- **Diskussionsthema:** Datenschutz / Datensicherheit.

# Technologien im Big Data-Umfeld



# Alle reden davon … Hadoop (V1.x) …

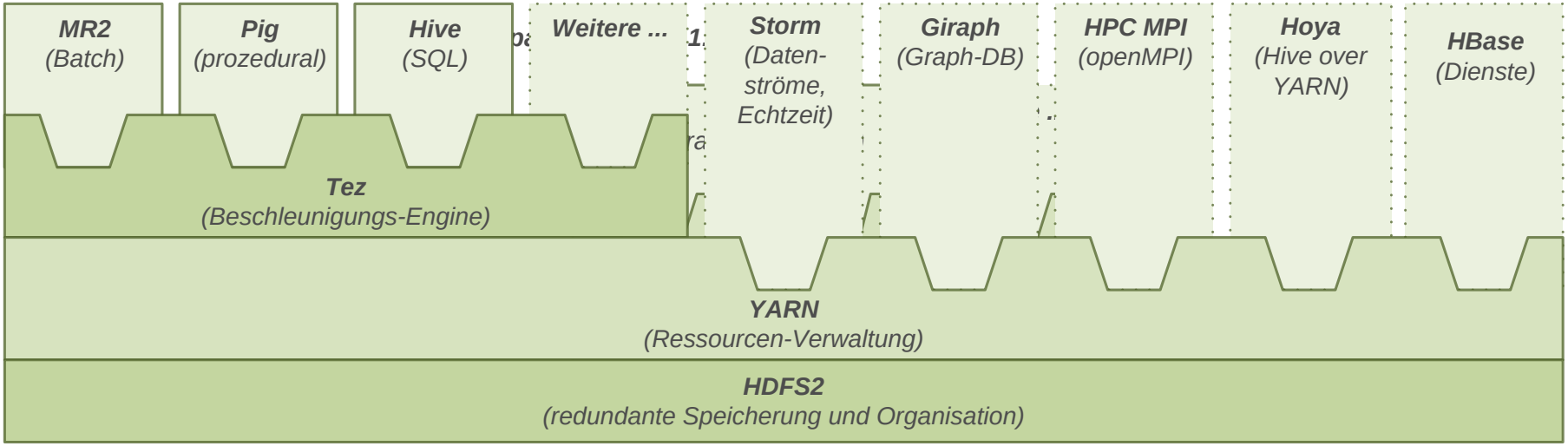



- **Apache Hadoop** ist eine Open Source Plattform für die Speicherung und die Verarbeitung von Daten
  - Skalierbar
  - Fehlertolerant
  - Verteilt
  - In Java geschrieben.
- **Speicherung und Auswertung** jeglicher Arten von Daten
  - strukturiert / unstrukturiert
  - Nicht an ein bestimmtes Schema gebunden.
- **Standardhardware** mit annähernd linearer Skalierung über n Nodes.

# Hadoop … zum Zweiten

FUJITSU

***Apache Hadoop 2.x ...***




Quelle: CT[2]

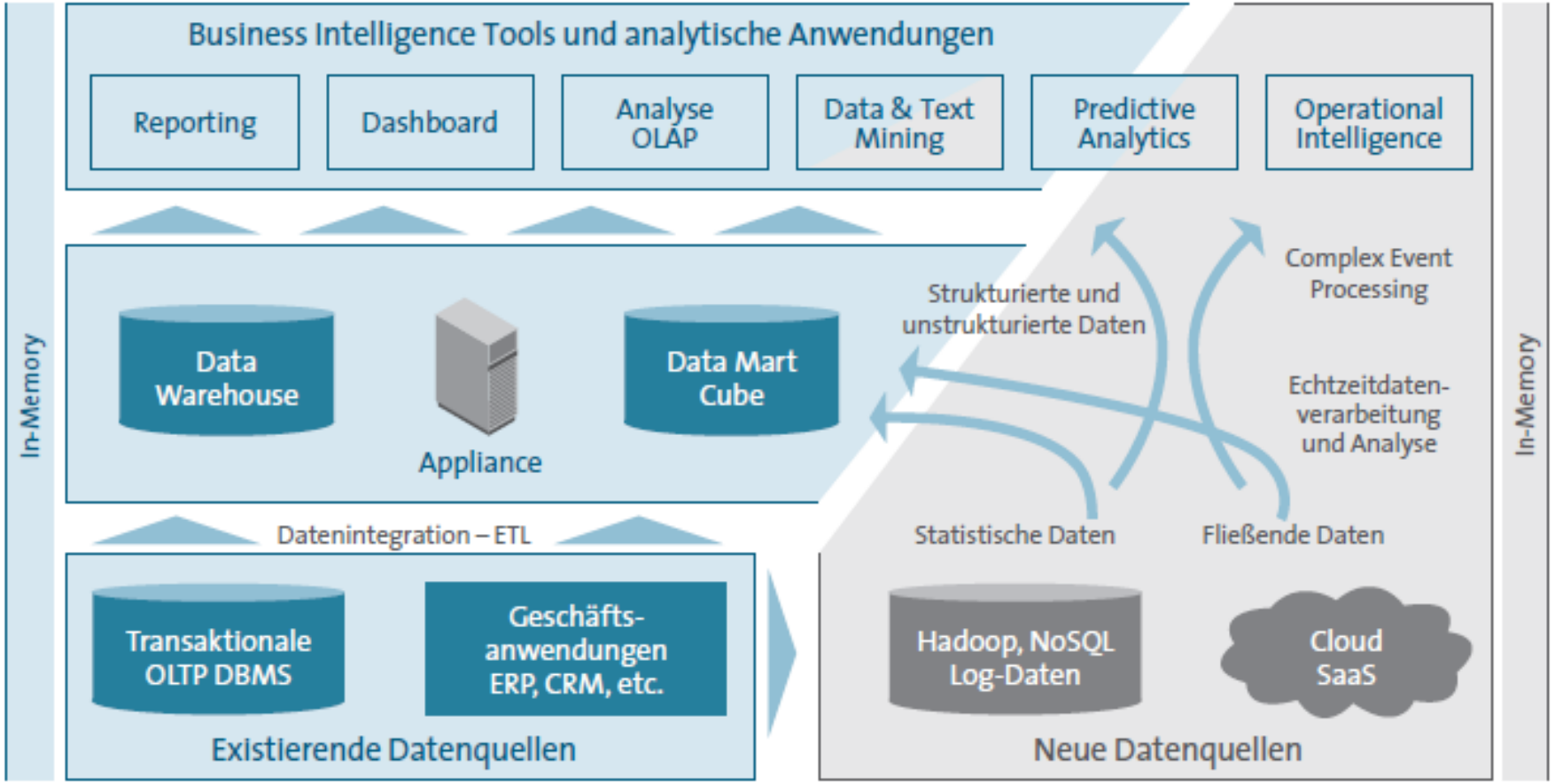
# Infrastrukturnahe Lösungen

FUJITSU

- In-Memory-Datenbanken (IMDB).
- Plattenspeicher mit In-Memory-Data-Grids (IMDG).
- In-Memory-Cache-Lösungen unter Einbeziehung von OpenSource (Ehcache mit den daraus resultierenden Einsatzbeschränkungen und Abhängigkeiten).
- Klassischer Ersatz bestehender Speicherlösungen (ggfs. unter Einsatz von SSD‘s … „kostenintensive vs. intelligenzintensiver Lösung“).

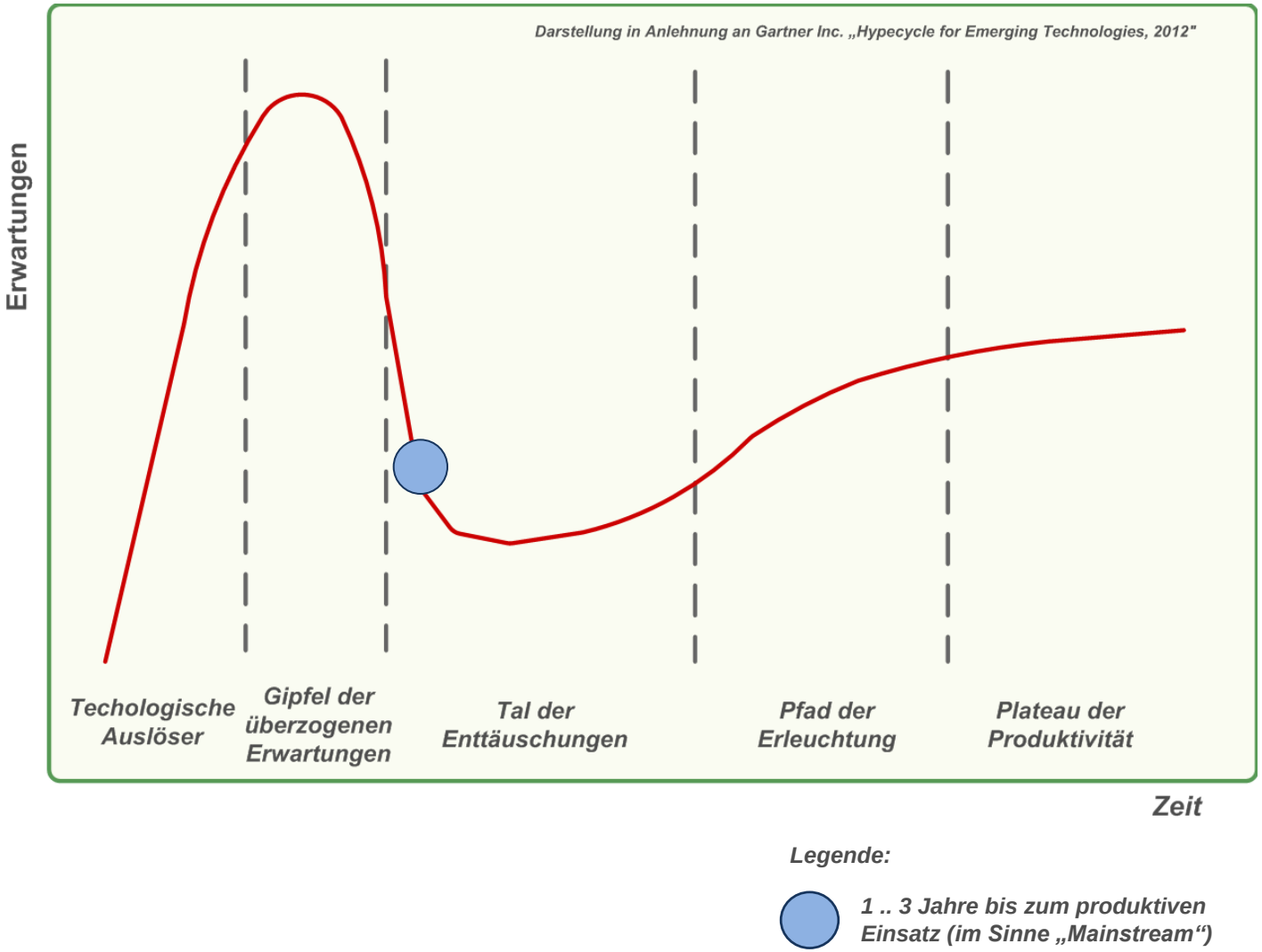
- Ausrichtung in der Regel auf Verringerung der Latenzen im Datenbereitstellungsprozess (HadoopFileSystem-Layer)!

# Wohin geht die Reise?




Quelle: BITKOM [2]

# Wo stehen wir heute?



# Wofür müssen Lösungen geliefert werden?

FUJITSU

- **Schnittstellen und Standards** („as a service“).
- **Nutzungsmodelle**.
- Sicherstellung der Einhaltung der **Datenschutzaspekte** über den gesamten Lifecycle der involvierten Prozesse.
- **Anwenderinteraktion** (EasyTo Use, SelfService).
- **Skalierbarkeit** über alle Prozessebenen (ja, es ist auch ein „Large Data“-Problem!).
- Unterstützende **Middleware**.
- **Big Data liefert Kerntechnologien**, Nutzungsszenarien sind der Kreativität der Marktbeteiligten geschuldet!

# Technologische Spannungsfelder

FUJITSU

- **Polystrukturierte Daten** aus den verschiedensten Quellen
- **Datenschutz /Zugriffsschutz** auf technischer Ebene
- Datenmanagement
- **Lifecycle der Daten** wird immer kürzer und wirkt sich verschärfend auf die Datenflut aus
- Lernen mit **Unschärfen** aus einzelnen Datenquellen zu leben, Validierung über mehrere Ebenen
  - Nachvollziehbare Analyse und Bewertung von Risiken
  - Bewertung von Datenquellen
  - Datengenauigkeit vs. Datenqualität (was ist „hinreichend“)
  - Werthaltigkeit kommt aus dem Zusammenhang und nicht aus dem Detail
- Zeitnahe Aufbereitung **entscheidungsfähiger Datenstrukturen**

# Gesellschaftliche Aspekte

FUJITSU

- These: Big Data ist von Nutzen für Wirtschaft und Gesellschaft, kann sich aber unter der Kontrolle von „Big Companies“ und „Big Government“ in das Gegenteil verkehren
- Stichwort: Wem gehören die Daten und wer hat Zugriff?
- These: Big Data-Projekte stellen Ergebnisse meist nur ihren Betreibern zur Verfügung
- Stichwort: Volkszählungsurteil von 1983
- Stichwort: Statistikgeheimnis, Re-Identifizierungsverbot (BStatG)
- Stichwort: Amtliche Statistik vs. BigData
- Stichwort: Industrie 4.0

# Big Data im Geo-Umfeld …

FUJITSU

- **Visualisierung** von raumbezogenen Daten mit zusätzlichen Kontexten.
- **Verknüpfungen** von relativ statischen Bestandsdaten mit dynamischen Daten in Echtzeit mit dem Raumbezug als Ordnungselement.
- **Schnelle Visualisierung** von bis dato nur als reine Zahlen verfügbares Datenmaterial (u. a. Sensorik).
- Einbeziehung **nutzergenerierter Geoinformationen** (Crowd Scouring).
- **Qualifizierung und Validierung der Datenqualität** (Hinreichende Verprobung von Plausibiltäten in Echtzeit über räumliche Strukturierungen).
- **Zielgruppenorientierte Bereitstellung** von Geodaten.

# Der Blick über den Tellerrand…

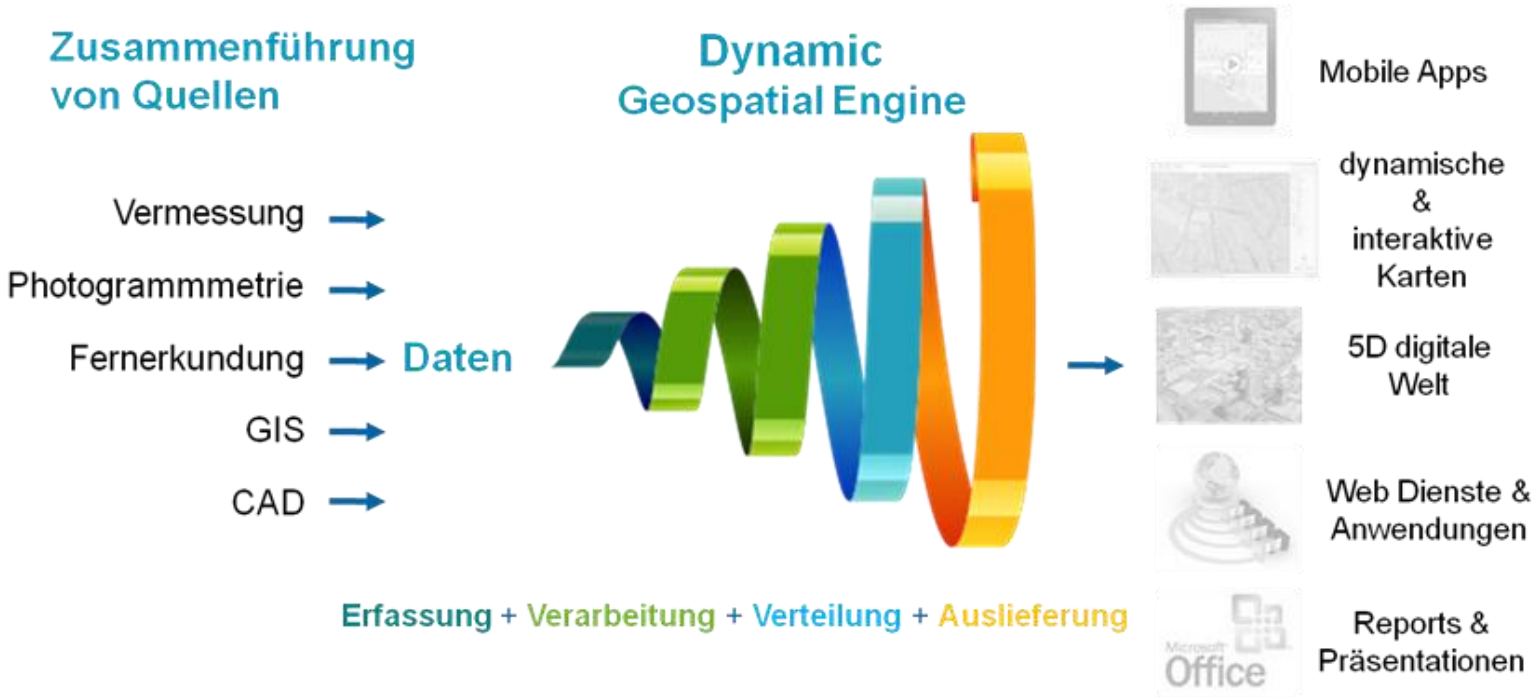
FUJITSU

- **Themen in den Steuerverwaltungen**
  - ELSTER – Teile von Hadoop seit 2012 im Einsatz
  - Finanztransaktionssteuer im Umfeld von „high-frequency trading“
  - Management von RZ-Infrastrukturen bei komplexer Vernetzung auf Verfahrensebene
  - Technologische Lösungen zur Absicherung von Koexistenzphasen von „bestehenden Verfahren“ auf Mainframesystemen zu neuen bzw. migrierten Verfahren auf Linux-Systemen
  - Kontinuitätsabsicherung
  - Automatisierte Veranlagung
  - Data Warehouse
  - Steuerfandung
  - Ankauf von Daten

# Herausforderungen … nicht nur im Geo-Umfeld!

- Neue Services aus dem Big Data-Umfeld sind mehr als nur die Bereitstellung von noch mehr Daten!

  Die anwachsende Datenflut stellt den Anwender vor zunehmende Probleme … qualifizierte und bewertete Daten sind gefragt (Aggregation des Raumes).



# Der Blick zum Anfang …
# …es ist jetzt vierzehn Uhr


Quelle: http://commons.wikimedia.org/wiki/File:4.29.11TimesSquareByLuigiNovi3.jpg

# Fragen? Danke!

FUJITSU

# Unterlagen

FUJITSU


(1) FUJITSU: Lösungsansätze Big Data
http://globalsp.ts.fujitsu.com/dmsp/Publications/public/wp-bigdata-solution-approaches-de.pdf

(2) BITKOM: Leitfäden Big Data
2012: https://www.bitkom.org/files/documents/BITKOM_LF_big_data_2012_online(1).pdf
2013: http://www.bitkom.org/files/documents/LF_big_data2013_web.pdf
2014: http://www.bitkom.org/files/documents/BITKOM_Leitfaden_Big-Data-Technologien-Wissen_fuer_Entscheider_Febr_2014.pdf

(3) Apache-Projekte:
http://hadoop.apache.org/
http://hbase.apache.org/
http://zookeeper.apache.org/
http://pig.apache.org/
http://hive.apache.org/
http://oozie.apache.org/
http://mahout.apache.org/
http://sqoop.apache.org/
http://flume.apache.org/