Big Data organisieren

Schritte zum Competence Center

Unter Big Data werden viele verschiedene technologische Facetten zusammengefasst. Die Meinungen sind gespalten, wenn es um eine genaue Deutung des Begriffs geht. Viele verstehen Big Data als eine Erweiterung von BI und deskriptiver Analytik um unstrukturierte Textanalysen oder einfach nur als Speicherung sehr großer Datenbestände.

Die anderen sehen in Big Data mehr als nur die, in der BI-Welt verbreitete, deskriptive Gewinnung von Erkenntnissen. Sie meinen, Big Data erweitere das klassische BI um Advanced Analytics [Gartner17] und ermögliche das Abbilden neuer Szenarios zur Datenbewirtschaftung auf Basis von bis dato unzugänglichen Datenquellen.

Beide Deutungen sind im Grunde genommen nicht falsch: In der Praxis erleben wir Big Data als technologische Weiterentwicklung bekannter Verfahren innerhalb eines breiten Datenuniversums. Aus dieser Sichtweise entstehen neue Anforderungen an eine Aufbauorganisation, die nun als Teil einer Big-Data-Strategie definiert werden kann. So unterliegen Big-Data-Lösungen einer wesentlich höheren Dynamik als kommerzielle Standardsoftware. Hinzu kommt, dass die Unternehmensorganisation bei Big Data besonders anpassungsfähig bleiben muss und somit in der Softwareentwicklung häufig agile Herangehensweisen Anwendung finden.

Neue Entwicklungsparadigmen und neue Analysemöglichkeiten benötigen einen veränderten, flexiblen Entwicklungsprozess, der eine explorative Vorgehensweise fördert. Neue Technologien, fordern neue Rollenprofile. All diese Faktoren sind wichtig, wenn Unternehmen Big Data als Wettbewerbsvorteil nutzen wollen.

WEB-TIPP:

www.opitz-consulting.com

„Alte“ Prozesse im neuen Gewand

Data Goverance, Ethik und Sandboxing sind fest etablierte Begriffe, es ist jedoch sinnvoll diese im Kontext von Big Data aus einem neuen Blickwinkel zu betrachten:

- **Data Governance wird zu Big Data Governance:** Die Data Governance sollte, entsprechend eingesetzter Datenquellen und verfügbarer Analysemöglichkeiten, an die gesetzlichen Regelungen angepasst werden. Je nach Vorhabengröße und Vielfältigkeit der eingesetzten Datenquellen und den damit einhergehenden potenziellen Sicherheitsrisiken, sollte die Einführung einer Big Data Governance in Betracht gezogen werden. [BCJMPRS14]

- **Es geht nicht ohne Ethik:** Ethische Aspekte werden bei Big-Data-Vorhaben immer wichtiger. Durch die analytischen Möglichkeiten im Big-Data-Kontext bekommen Unternehmen eine neue Sicht auf die Zusammenhänge in den Daten. Der Umgang mit diesen Erkenntnissen sollte aus der ethischen Perspektive genau überlegt sein. Einige Unternehmen haben bereits eine Corporate Social Responsibility (CSR) als Leitlinie integriert. [DP12]

- **Fail early, fail often:** In der PoC-oder Prototyp-Phase sind Big-Data-Vorhaben oft hypothesengetrieben. Sinnvoller erscheint in dieser Phase ein exploratives Vorgehen, das einem Lean-Startup-Ansatz entspricht [Ries 2014]. Im explorativen Ansatz ist der wirtschaftliche Nutzen nicht immer bekannt. Er wird erst im Zuge eines iterativen Vorgehens ermittelt, wenn die aufgestellte Hypothese auf ihre Plausibilität hin überprüft wird. Ein wichtiger Aspekt ist hier das Motto „Fail early“. In der Praxis bedeutet dies, dass das Ergebnis einer Datenanalyse oder einer statistischen Modellierung die Erwartungshaltung nicht erfüllt. Aber auch ein negatives Ergebnis kann eine neue Hypothese hervorbringen, die weiter iterativ getestet wird. [DFFST16]

- **Sandboxing:** Erfahrungen zeigen, dass Unternehmen, die mit anonymisierten oder generischen Daten experimentiert haben, manchmal erst bei der Überführung der Modelle in die Produktivumgebung feststellen, dass das trainierte Modell, da es nicht auf echten Daten aufbaute, nicht die erwarteten Ergebnisse auf Produktivdaten liefert. Vorteilhaft kann es daher sein, den Sandboxprozess anzupassen und generell eine mehrstufige Datentrennung anzustreben. Diese lässt zum einen dem Data Scientist seine Freiheit für die Einbindung von 3rd-Party-Data und zum anderen stellt sie sicher, dass Replikationen von produktivnahen Daten beim Trainieren eines Modells zur Verfügung stehen. Der Unterschied zwischen einer Thesenerprobung innerhalb einer Sandboxumgebung (Data Lab) und innerhalb einer permanenten Datenbewirtschaftung (Data Factory) auf Basis von erprobten Verfahren, kann wie folgt zusammengefasst werden:

Versionsmanagement auch im Fachbereich

Befindet sich ein Data Scientist in einem Fachbereich, dann kann es ratsam sein, auch im Fachbereich ein Versionsmanagement einzuführen. So hat der Data Scientist seine Freiheit: Er kann mit Replikationen von Produktivdaten arbeiten und gleichzeitig beliebige Daten in die Data-Lab-Umgebung importieren, ohne befürchten zu müssen, dass er später nicht mehr auf eine frühere Version seines Codes zurückgreifen kann. Das Versionsmanagement hilft auch dem Big Data Developer, wenn es in der nächsten Phase darum geht, das Verfahren aus dem Data Lab in die Produktion (Data Factory) zu überführen.

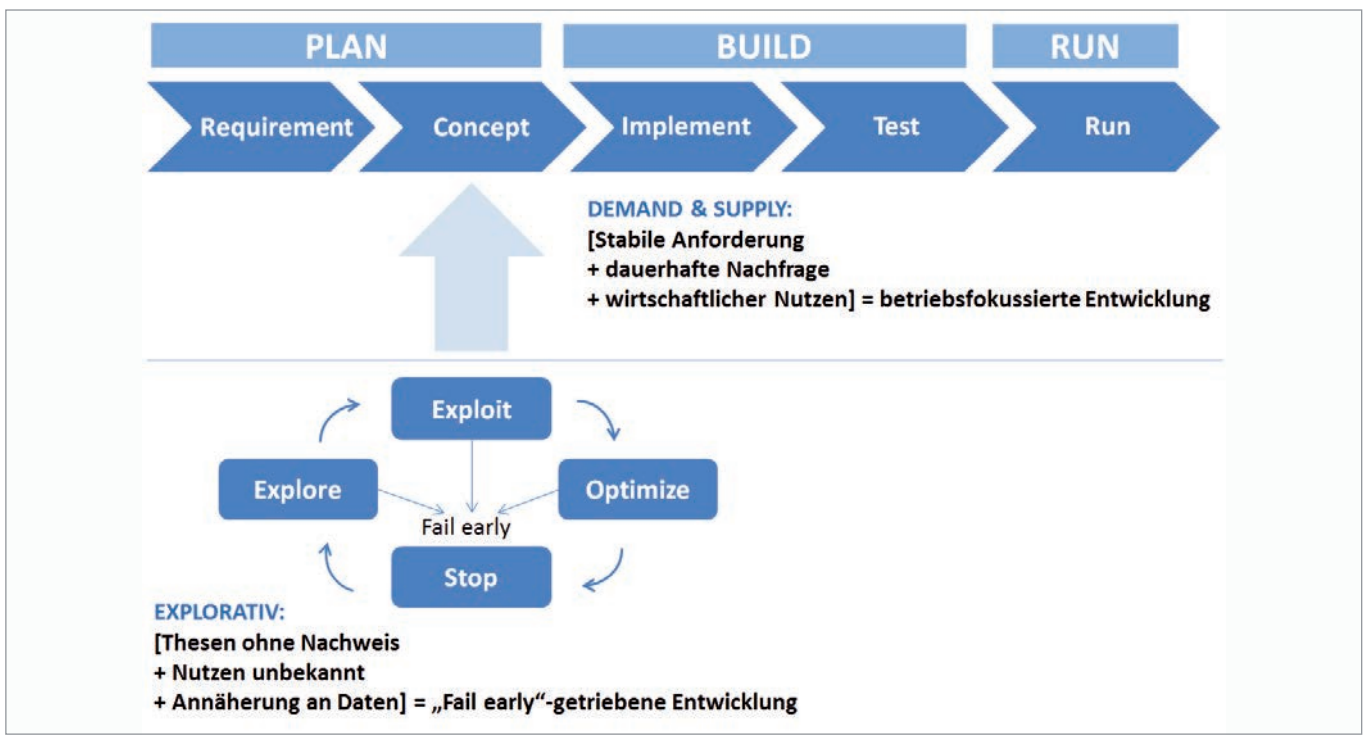

Bild 1: Demand & Supply vs. explorative Vorgehensweise.

Welche Änderungen kommen auf die Rollen zu?

Ähnlich zu den oben genannten Prozessen müssen auch die Skills der Rollenprofile an die neuen technologischen Anforderungen angepasst oder sogar neue Rollenprofile definiert werden.

Erweiterung bestehender Rollen

Die Rolle des Data Scientist ist bereits aus der BI-Welt bekannt. Sie wird lediglich um den Umgang mit einem Hadoop Frontend (etwa Hue) erweitert sowie um das Grundverständnis von Datenhaltung auf HDFS und je nach Anwendungsfall auch um den Umgang mit unstrukturierten Daten (etwa Text). Falls der Data Scientist R oder Python noch nicht anwendet, kommen diese noch hinzu. In vielen Unternehmen kann es sinnvoll sein, mit dem hohen Datenaufkommen durch neue Datenquellen, die Datenaufbereitungstätigkeit von den sonstigen Aufgaben eines Data Scientists zu entkoppeln. Diese Aufgabe kann je nach Ausmaß entweder einer separaten Rolle oder einer bestehenden Rolle mit weniger Auslastung zugeordnet werden. Der Data Scientist kann sich dann besser auf seine Kernaufgabe konzentrieren: Die Gewinnung von Insights durch die Überprüfung von Hypothesen.

Mögliche neue Rollen

- **Big Data Architect:** Diese Rolle erfordert eine tiefe Kenntnis in verfügbaren Big-Data-Technologien und aktuellen Trends in diesem Gebiet. Im Optimalfall hat der Big Data Architect auch Erfahrungen in BI-Architektur.

- **Big Data Developer:** Dieser monetisiert die gewonnenen Insights, indem er die verprobten Algorithmen und Modelle aus dem Data Science in die Produktivumgebung überführt (z. B. eine Scala-Implementierung eines Algorithmus, den ein Data Scientist in R geschrieben hat). Der Big Data Developer arbeitet eng mit dem Data Scientist zusammen, wenn die Verfahren aus dem Data Lab in die Data Factory überführt werden.

- **Big Data Integrator:** Er ist für die Implementierung von performanten, systemübergreifenden Data Pipelines zuständig. Von der Anbindung einer neuen Datenquelle, über Datenintegration bis hin zur Verwendung in analytischen Modellen und Operationalisierung ist somit jeder Teilbereich vertreten.

- **Big Data System Engineer:** Dieser ist für die Installation und den Betrieb des Hadoop Clusters zuständig, überwacht die Prozesse und administriert Hardware und Applikationen.

Die richtige Form der Aufbauorganisation

Viele Unternehmen müssen bei der Einführung von Big-Data-Technologien eine Hürde überwinden. Oft sind ihre Funktionsbereiche nach einer rigiden Plan-Build-Run-Form ausgerichtet [Zarnekow07]. Durch mehrere Führungskräfte, viele Schnittstellen und dadurch bedingte hohen Durchlaufzeiten entstehen möglicherweise Interessenskonflikte. Diese setzen die Unternehmen unter den Druck, für den Betrieb und die Weiterentwicklung von Produkten auf Basis von Big Data eine andere, agile Aufbauorganisation schaffen. Die folgenden Aspekte helfen ihnen, in diesem Fall einen schnellen Einstieg in Big Data zu schaffen:

Neue Technologien, fordern neue Rollenprofile. All diese Faktoren sind wichtig, wenn Unternehmen Big Data als Wettbewerbsvorteil nutzen wollen."

Dimitri Gross, Senior Consultant
Opitz Consulting AG

- Eine möglichst schnittstellenarme Möglichkeit zur Zusammenarbeit mit dem Fachbereich, dessen IT-Affinität durch das Trendthema Digitalisierung forciert wird.
- Niedrige Organisationshürden und eine höhere Flexibilität, um den Einstieg in eine neue Technologie mit vielen Unbekannten optimal zu schaffen.

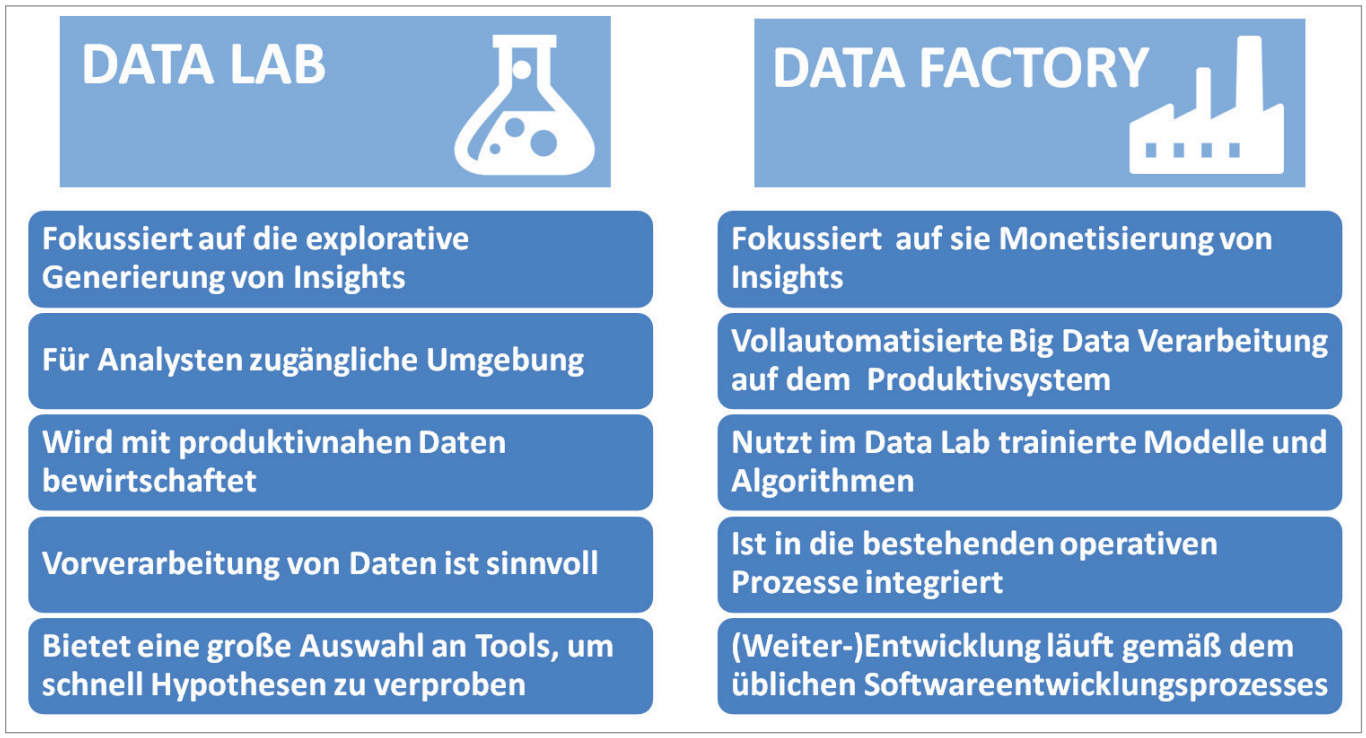

Bild 2: Data Lab vs. Data Factory.

Viele Formen der Aufbauorganisation, die aus der Organisationslehre bekannt sind, kommen dabei in Frage. Von Linien- über Stab- und Matrix-Modellen bis zur funktionalen Organisation in einer virtuellen Variante oder als eigenständige Organisation ist alles denkbar. Die Wahl der passenden Aufbauorganisation richtet sich individuell nach der Unternehmensgröße, der strategischen Bedeutung des Big-Data-Vorhabens sowie den geplanten Anwendungsfällen und der aktuell gültigen Aufbauorganisation. [DFFST16]

Viele Unternehmen gehen den ersten Schritt in Richtung eines virtuellen Projektteams mit einer fest definierten Arbeitszeitkapazität (z. B. 60 % der Arbeitszeit wird für PoC reserviert). Damit erreichen sie bereits einen großen Fortschritt in Richtung Agilität.

Was ist noch zu beachten?

Unternehmen, die Big Data erfolgreich umsetzen möchten, sollten erfahrungsgemäß noch diese Punkte im Auge behalten und mittelfristig angehen:

- Schaffung eines permanenten Wissenstransfers in einer internen Community zur Förderung des Know-how-Aufbaus.
- Durchführung von Retrospektiven zu jeder Projektphase in regelmäßigen Abständen.
- Enge Zusammenarbeit mit Fachbereichen. Die fachlichen Anforderungen sollen persönlich transportiert, stetig geprüft und nachgehalten werden.
- Schlankes Anforderungsmanagement mit möglichst kurzen Zyklen und ständige Priorisierung von Aufgaben.
- Zusammenarbeit mit Universitäten, um das Recruiting zu optimieren.

Fazit

Big Data ist kein Fremdwort mehr und hat bereits viele neue Geschäftsmodelle hervorgebracht. Es schafft Raum für neue Ideen und Möglichkeiten. Mit der richtigen Strategie, gelingt auch der erste Schritt in die neue Technologieära. Unternehmen können ihr Risiko minimieren, indem sie im Rahmen einer iterativ angelegten Erprobungsphase zunächst die Wichtigkeit für die Gesamtunternehmung bewerten sowie Schwachstellen und Risiken ermitteln.

Dimitri Gross

LITERATUR:

[DFFST16] Big Data Ein Überblick, Carsten Dittmar, Carsten Felden, Ralf Finger, Rolf Scheuch, Lars Tams, dpunkt.verlag GmbH, 2016

[Zarnekow 07] Produktionsmanagement von IT-Dienstleistungen: Grundlagen, Aufgaben und Prozesse, Ruediger Zarnekow, Springer Science & Business Media, 09.01.2007 - 293 Seiten

[DP12] Ethics of Big Data, Kord Davis, Doug Patterson ,„O'Reilly Media, Inc.", 30.06.2012

[BCJMPRS14] Information Governance Principles and Practices for a Big Data Landscape, Chuck Ballard, Cindy Compert, Tom Jesionowski, Ivan Milman, Bill Plants, Barry Rosen, Harald Smith, IBM Redbooks, 31.03.2014 - 280 Seiten

[Ries 2014] Ries, E.: The lean startup: how today's entrepreneurs use, continuous innovation to create radically successful businesses. Crown Publishing, 2011, 2014.

[Gartner 2017] Gartner IT Glossary / Advanced Analytics, http://www.gartner.com/it-glossary/advanced-analytics/ (abgerufen am 18.08.216)