Hochverfügbarkeit für Manager
- klüger entscheiden

Daniel Steiger
Trivadis AG
Zürich (Schweiz)

Schlüsselworte:

Hochverfügbarkeit, High Availability, HA, Business Impact, Datenbankinfrastruktur

Hochverfügbarkeit ist ein entscheidender Baustein Ihres Geschäftserfolges

Die Geschäftsmodelle erfolgreicher Unternehmen basieren zunehmend auf der Verfügbarkeit der Unternehmensdaten: no data – no business – no success. Als Folge davon lautet das Credo: „Always On". Dies betrifft natürlich besonders die IT als Ganzes und die IT-Infrastruktur im Speziellen. Systemausfälle oder Systemunterbrüche - geplante oder ungeplante - behindern die Geschäftsabläufe, oder führen gar zu einem Geschäftsunterbruch, was in den meisten Fällen mit Kosten oder Imageverlust verbunden ist. Doch wie hoch sind die effektiven Businesskosten und was sind die Auswirkungen eines Systemausfalls? Welche Geschäftsprozesse sind betroffen? Wie lange dauert es bis der Service wiederhergestellt ist? Und was kann unternommen werden um die Datenverfügbarkeit zu erhöhen und Störungen zu vermeiden?
Diese und weitere Fragen müssen auf dem Weg zur optimalen Lösung, welcher durch die grosse Auswahl von zunehmend komplexeren Technologien und Systemen auch nicht einfacher wird, beantwortet werden. Insgesamt also ein schwieriges Terrain für Entscheidungsträger, die vor allem eines wollen: Gewissheit und Sicherheit in ihren Entscheidungen – und Technik die begeistert!

Wir zeigen Ihnen auf, wie Sie die damit verbundene Komplexität meistern können

Die mit Infrastrukturentscheiden verbundene Komplexität können wir nicht aus der Welt schaffen. Aber wir können anhand eines Leitfadens aufzeigen, wie Sie trotzdem bessere und damit klügere Entscheidungen fällen können. Eine wichtige Voraussetzung dazu ist, dass Sie die Geschäftsanforderungen wie Zuverlässigkeit, Verfügbarkeit und Performanz methodisch klären und in Infrastrukturanforderungen übersetzen können. Kenntnisse der essentiellen Verfügbarkeitsbegriffe, Blueprints und Technologietrends helfen, eine optimale Lösung zu finden.

Leitfaden für klügere Entscheidungen

Im Zentrum der Evaluation einer Hochverfügbarkeitslösung stehen die spezifischen Geschäftsanforderungen. Diese stehen über allen anderen Kriterien. Stimmt beispielsweise das Kosten-Nutzen-Verhältnis nicht, dann nützt die beste und neuste Technologie nichts.
Im Gegensatz zur Verfügbarkeitsdiskussion (welche oft eine Technische ist) liegt die geschäftsbezogene Risikobetrachtung den Entscheidungsträgern naturgemäss näher. Dazu eignet sich die sog. Business Impact Analysis – ein schrittweises Vorgehen, mit welcher die geschäftsspezifischen

Auswirkungen abgeschätzt werden. Eine umfassende Business Impact Analyse adressiert folgende Themen:

- Identifikation der kritischen Geschäftsprozesse
- Ermittlung des quantifizierbaren Risikos aufgrund geplanter und ungeplanter Ausfälle von IT-Systemen auf die Geschäftsprozesse
- Beschreibung der Auswirkung der Ausfälle

Als Leitfaden für die Spezifikation der businessspezifischen Verfügbarkeitsanforderungen hat sich folgendes Vorgehen[1] bewährt:

1. Analyse der Auswirkung auf das Geschäft und die kritischen Geschäftsprozesse
2. Analyse der Ausfallkosten (Cost of Downtime)
3. Bestimmung der Recovery Time Objective (RTO)
4. Bestimmung des Recovery Point Objective (RPO)
5. Festlegen des Manageability-Ziels (Komplexitätstoleranz der Organisation)
6. Berechnung der Total Cost of Ownership (TCO) und Return on Investment (ROI)

Im Rahmen der „Business Impact Analyse" müssen diejenigen Fragen geklärt werden, die wesentlichen Einfluss auf die Systemarchitektur haben:

- Wie hoch sind die Kosten eines Systemausfalls?
- Welche Daten müssen gegen Verlust geschützt werden und wie gross ist die Datenmenge?
- Gegen welche Art von Katastrophen und Ausfälle müssen Sie sich schützen?
- Wie lange kann ihr Geschäft ohne System überleben?
- Wie lange ist das Zeitfenster, während dem das System nicht verfügbar sein kann?
- Ist ein bestimmter Datenverlust tolerierbar?
- Wie gross ist die maximale Datenmenge, die verloren gehen kann?

Die obigen Anforderungen bilden die Basis für die Systemarchitektur. Deren Ausarbeitung liegt typischerweise in der Verantwortung des IT-Managements. Unter Berücksichtigung der strategischen und technischen Rahmenbedingungen haben sich folgende Leitlinien bewährt:

7. Technische Umsetzung entlang den RASP-Kriterien (Reliability, Availability, Scaleability und Performance)
8. Reduktion der Komplexität in der Implementierung und im Betrieb; d.h. standartisieren, automatisieren, dokumentieren und trainieren
9. Ggf. Begleitung durch einen Trusted Partner während des gesamten Lifecycles

HA-Essentials

Im Folgenden erläutern wir die essentiellen Begriffe aus der Welt der „Hochverfügbarkeit". Betrachtet man die Begriffe aus zwei Perspektiven, nämlich aus Businesssicht und aus technischer Sicht wird klar, dass in deren Interpretation immer eine bestimmte Unschärfe enthalten ist. Relevant und schlussendlich entscheidend sind auf jeden Fall die Geschäftsanforderungen, welche so differenziert wie möglich beschrieben sein sollten.

[1] In Anlehnung an „Analysis Framework for Determining High Availability Requirements" aus Oracle High Availability Overview 11g Release 2, E10804-01

Zuverlässigkeit und Verfügbarkeit
Oracle schlägt vor die Anforderungen entlang den RASP-Kriterien (Reliability, Availability, Scaleability und Performance) zu beschreiben. Zuverlässigkeit und Verfügbarkeit zählen dabei ohne Zweifel zu den wichtigsten architekturtreibenden Faktoren. Unter **Verlässlichkeit** (Reliability) versteht man die Wahrscheinlichkeit, dass ein bestimmtes System während einer bestimmten Zeit "störungsfrei" funktioniert. Herausragendes Beispiel eines Systems mit sehr hoher Verlässlichkeit sind die Verkehrsflugzeuge. Deren Verlässlichkeit liegt bei nahezu 100%[2]. Weit weniger verlässlich ist beispielsweise das mobile Telefonnetz: Gesprächsunterbrüche und Störungen sind an der Tagesordnung. Die **Verfügbarkeit** (Availability) bezeichnet die „Up-Time" eines Systems in %. Die Verfügbarkeit ist zwar eine wichtige Voraussetzung für die Daten- und Serviceverfügbarkeit, aber nur beschränkt aussagekräftig bezüglich der Servicequalität.

In Tabelle 1 sind die Ausfallzeiten für eine bestimmte Verfügbarkeit in % aufgelistet (bezüglich einer Gesamtzeit von 24 Stunden x 365 Tagen). Man bezeichnet die Verfügbarkeitsklassen auch als „n Nines", also beispielsweise „3-Nines" oder „4-Nines".

Availability %	Downtime per year	Downtime per month*	Downtime per week
98%	7.30 days	14.4 hours	3.36 hours
99%	3.65 days	7.20 hours	1.68 hours
99.5%	1.83 days	3.60 hours	50.4 min
99.9%	8.76 hours	43.2 min	10.1 min
99.99%	52.6 min	4.32 min	1.01 min
99.999%	5.26 min	25.9 s	6.05 s
99.9999%	31.5 s	2.59 s	0.605 s

Tabelle 1: Verfügbarkeitskennzahlen

Recovery Point Objective (RPO) und Recovery Time Objective (RTO)
RPO definiert das maximal zulässige Zeitfenster, aus dem Transaktionsdaten - aufgrund eines Disasters, resp. Systemunterbruchs - verloren gehen dürfen (Angabe in Sekunden, Minuten, oder Stunden). RTO definiert die maximal tolerierbare Serviceausfallzeit (Angabe in Sekunden, Minuten, Stunden oder Tagen) aufgrund eines Disasters, resp. Systemunterbruchs.

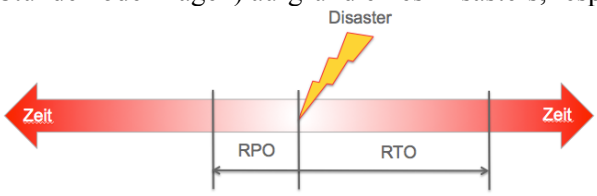

Abbildung 1: RPO und RTO

Der RPO-Begriff ist technisch betrachtet etwas „unscharf", weil er ein Zeitfenster und nicht eine exakt definierten Datenmenge beschreibt. An dieser Stelle muss man die Geschäftsprozesse und Datenflüsse gut kennen, um die Auswirkungen eines allfälligen Datenverlustes abschätzen zu können. RPO bestimmt wesentlich den Technologieeinsatz und stellt für den Systemarchitekten eine klare Vorgabe

[2] Faktoren die zu dieser hohen Verlässlichkeit führen sind u.a. standartisierte Checks und Wartungsintervalle, konsequentes Lifecyclemanagement, redundante Systeme, hochqualifiziertes Personal und eine ausgeprägte Sicherheitskultur.

dar. In der Praxis wird RPO = Null (No Data Loss) selten umgesetzt, weil die damit verbundenen Restriktionen zu einschneidend, resp. die Kosten zu hoch sind. In der Praxis wird üblicherweise „Near-Zero Data Loss“ („RPO fast Null“) angestrebt.

RTO beschreibt als „Business Continuity“-Kennzahl die „gefühlte Verfügbarkeit“ am treffendsten. Tendiert die Serviceaufallzeit gegen Null, d.h. die tolerierte Ausfallzeit liegt im Minuten-, oder gar Sekundenbereich, schnellen die Kosten für die technische Umsetzung überproportional in die Höhe (siehe Abbildung 2). Wie RPO, ist die RTO-Anforderung eine klare Vorgabe an die System-architektur.

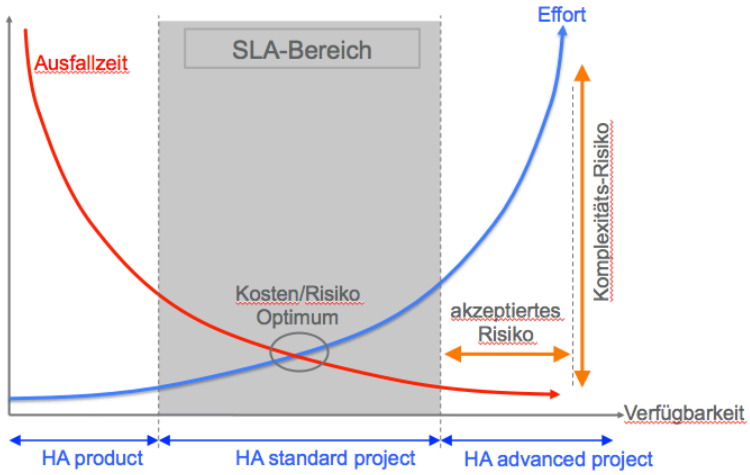

Abbildung 2: Verfügbarkeit vs. Kosten

Service- vs. Datenhochverfügbarkeit
Die beiden Begriffe Service- und Datenhochverfügbarkeit werden in der Praxis oft nicht klar unterschieden. **Daten-Hochverfügbarkeit** bedingt eine vollständige und unabhängige Vervielfältigung der Daten (Datenredundanz). Dies geschieht typischerweise mittels logischer oder physischer Spiegelung über zwei oder mehr Standorte hinweg. Oracle Data Guard ist eine bewährte Technologie um Daten-HA sicherzustellen. **Service-Hochverfügbarkeit** auf der anderen Seite, bedingt eine Vervielfältigung der Services auf mehrere Server oder Standorte, wobei die Daten nicht zwingend multipliziert werden müssen. Oracle Real Application Cluster (RAC) ist eine bewährte Service-HA-Lösung.

Geplante vs. ungeplante Ausfälle
Normalerweise liegt der Fokus einer HA-Lösung auf der Risikoverminderung aufgrund ungeplanter Ausfälle wie Hardwareausfall, Rechenzentrumsausfall, Datenkorruption oder Operatorfehler. Doch auch geplante Ausfälle, wie System-, Daten- und Applikationsänderungen verursachen Kosten, die oft unterschätzt, oder einfach „ausgeblendet“ werden. Die Reduktion geplanter Ausfälle sollte deshalb genauso in die HA-Lösung einfliessen, wie die Massnahmen gegen ungeplante Ausfälle.

HA-Blueprints und Technologietrends

Im Wesentlichen verfügt Oracle über vier Basisarchitekturen, die auf unterschiedlichste Art und Weise implementiert werden können (physisch, virtuell, verteilt über Lokationen, etc.):

- Single-Instanz-Datenbank auf Single-Server
- Single-Instanz-Datenbank auf einem Failover-Cluster

- Multi-Instanz-Datenbank auf einem Server-Cluster (RAC)
- Standby-Datenbank (Data Guard, für Single- oder Multi-Instanz-Datenbanken)

RAC und Data Guard können zur sog. Maximum Availability Architektur (MAA) kombiniert werden. In Tabelle 2 sind die Charakteristiken der Basisarchitekturen beschrieben.

Architektur	Charakteristik
Single Instance	Keine, resp. teilweise Redundanz. Systemausfall auch beim Ausfall einer Komponente. Scale-Up begrenzt möglich (Einsatz von Instance Caging möglich ab 11gR2).
Failover Cluster (FOC)	Knotenredundanz. Service-Failover bei Knotenausfall. Automatischer Failover mit 3rd-Party Clustermanager, mit Oracle Clusterware,oder mit Oracle Failsafe auf Windows. Kein Disasterschutz
Real Application Cluster (RAC)	Skalierungslösung (Scale-Out) und Service-Hochverfügbarkeit. Kein Disasterschutz (ausser mit Streched-Cluster und Host-Based-Mirroring).
Data Guard (DG)	Schutz vor Disaster auf Basis einer aktiven Standby-Umgebung (Standby-DB). Automatischer Failover bei Serviceausfall auf Primärseite (mittels Fast-Start-Failover und Observer). Verschiedene Protection-Levels von Max-Performance bis Max-Availability.
MAA	Maximum Availability Architecture: Kombination vom RAC und Data Guard. Skalierfähige Daten- und Service-Hochverfügbarkeitslösung

Tabelle 2: Oracle Basisarchitekturen

Zusätzliche Verfügbarkeits-Features und-Produkte ergänzen die Basisarchitekturen und erlauben eine auf die spezifischen Anforderungen ausgerichtete Implementierung. Beispiele sind die Oracle Grid Infrastruktur (Clusterware und ASM), welche im Bereich Skalierung, Konsolidierung und Failover sowohl für RAC, wie auch für Failover-Cluster interessante Möglichkeiten[3] eröffnet.

Technologietrends

Hochverfügbare Systeme sind komplexe Systeme – im Aufbau wie im Betrieb. Die Reduktion dieser inhärenten Komplexität muss das Ziel jedes HA-Projektes sein. Die sogenannten Engineered-Systems wie Oracle Exdata, Exalogic oder Database Appliance bieten hier Abhilfe. Sie unterscheiden sich gegenüber konventionellen Systemen massgeblich. Engineered-Systems sind optimal auf ihre spezifische Aufgabe abgestimmte Komplettsysteme – Storage, Network, Server und Software. Der Engineeringaufwand, und das damit verbundene Risiko, schrumpft gegen Null und auch der Betrieb ist Dank der hohen Integration und dem Herstellersupport einfacher. Engineered-Systems sind daher speziell für hochverfügbare Datenbbanken eine interessante Alternative zu den klassischen Infrastrukturen. Auch Cloud-Computing entwickelt sich je länger je mehr als Basis für hochverfügbare Datenbankinfrastrukturen. Finanziell interessant sind (Public-)Cloudlösungen vor allem wegen dem auf den operativen Kosten beruhendem Finanzierungsmodell (Opex), im Gegensatz zum kapitalintensiven Capex-Modell. Am weitesten verbreitet ist bis anhin trotzdem der Einsatz von Private Cloud-Lösungen. Public Cloud-Lösungen und Hybrid Cloud-Services werden für HA-Datenbanklösungen bislang nur von wenigen Unternehmen genutzt. Wichtig bei Cloudlösungen: es müssen dieselben Anforderungen und Prinzipien zur Anwendung kommen, wie bei Non-Cloud-Infrastrukturen.

[3] Beispielsweise RAC One Node: Failover mit OMotion erlaubt flexibles Capacitymanagement, oder CRS only: Failoverlösung, bedingt lediglich die Standard Edition, ohne RAC.

Hochverfügbarkeit ist kein Produkt – HA ist ein Projekt

Eine optimal auf die Geschäftsanforderungen ausgerichtete Verfügbarkeitslösung zu evaluieren ist eine Herausforderung, die nur durch die enge Zusammenarbeit von IT- und Businessmanager erfolgreich gemeistert werden kann[4]. In der Praxis bewegt man sich dabei immer im Spannungsfeld zwischen geschäftlichen Anforderungen und technischer Machbarkeit (resp. Kosten). Abbildung 2 zeigt diesen Zusammenhang auf. Die optimale Lösung liegt irgendwo im grau markierten Bereich („Kompromisszone", resp. SLA-Bereich). Eine HA-Lösung ist zudem immer ein Puzzle aus Technologie, Betriebskonzept und Lifecyclemanagement. Standartisierung, Automatisierung, Dokumentation und Mitarbeiterschulung sind nicht zu vernachlässigende Faktoren, die zur Komplexitätsreduktion und zur Datenverfügbarkeit beitragen. Und daran wird die HA-Lösung am Ende des Tages gemessen:
Erfolg = Verfügbarkeit der Unternehmensdaten sicherstellen, wann immer sie benötigt werden.

Fazit und Take-Aways

- Die Businessanforderungen stehen im Zentrum – untersuchen Sie die Auswirkungen von Systemausfällen auf den Geschäftsbetrieb gründlich und vollständig
- Hochverfügbarkeit ist Verhandlungssache – zwischen IT-Management und Businessmanagement
- Keep it simple – reduzieren Sie Komplexität wo möglich und sinnvoll
- Prüfen Sie den Einsatz von neuen Konzepten und Produkten wie Engineered-Systems oder Cloud-Lösungen
- Hochverfügbarkeit ist ein Projekt und ist kein Produkt das Sie „ab Stange" kaufen können

Literaturhinweise

- Oracle® Database High Availability Overview, 11g Release 2 (11.2), Part Number E17157-08
- Enterprise Data and the Cost of Downtime, 2012 IOUG Database Availability Survey

Kontaktadresse:

Daniel Steiger
Partner / Principal Consultant
Trivadis AG
Europastrasse 5
CH-8152 Glattbrugg (Schweiz)

Telefon: +41 (44) 808-70-20
Fax: +41 (44) 808-70-21
E-Mail daniel.steiger@trivadis.com
Internet: www.trivadis.com

[4] Man spricht in diesem Zusammenhang auch von einem „Enterprise Management Approach", d.h. die Geschäftsleitung, resp. die Businessmanager bestimmen und treiben aktiv Entscheidungen bezüglich der Datenverfügbarkeit.