Statistische Analyse von Performance-Daten

Nicholas Dille
Microsoft MVP für RDS

Wer ist Nicholas Dille?

- IT-Architekt bei der sepago GmbH
  - Strategieberatung
  - Technische Konzeption
- Kernkompetenzen
  - Zentralisierung
  - Anwendungsbereitstellung
  - Kapazitätsmanagement

- Microsoft MVP für Remote Desktop Services (RDS)
- Blog: http://blogs.sepago.de/nicholas

Warum Performance-Analyse?

- Monitoring
- Schwellwertüberwachung
- Impact von Anwendungskomponenten
- Impact von Benutzerzugriffen
- Etablieren einer Baseline
- Sizing von Systemen
- Analyse von Bottlenecks
- SLA-Überwachung
- Lastprognosen

Aspekte der Performance-Analyse

- Echtzeitalarmierung – der Verkehrspolizist
  - Kurzfristige Aussagen über den Zustand
  - Überwachung von Performance-Metriken
  - Bewertung durch Schwellwerte
  - Kurze bis keine Aufbewahrungsfrist

- Kapazitätsmanagement – die Verkehrsleitzentrale
  - Mittel- bis langfristige Aussagen
  - Überwachung von Schlüsselmetriken
  - Trend-Analyse dieser Daten
  - Erstellung von Prognosen
  - Mehrmonatige Aufbewahrungsfrist

Arbeitsschritte

- Sammeln
  - PerfMon ist immer vorhanden

- Analysieren
  - SQL, Excel

- Visualisieren
  - Reporting Services, Excel

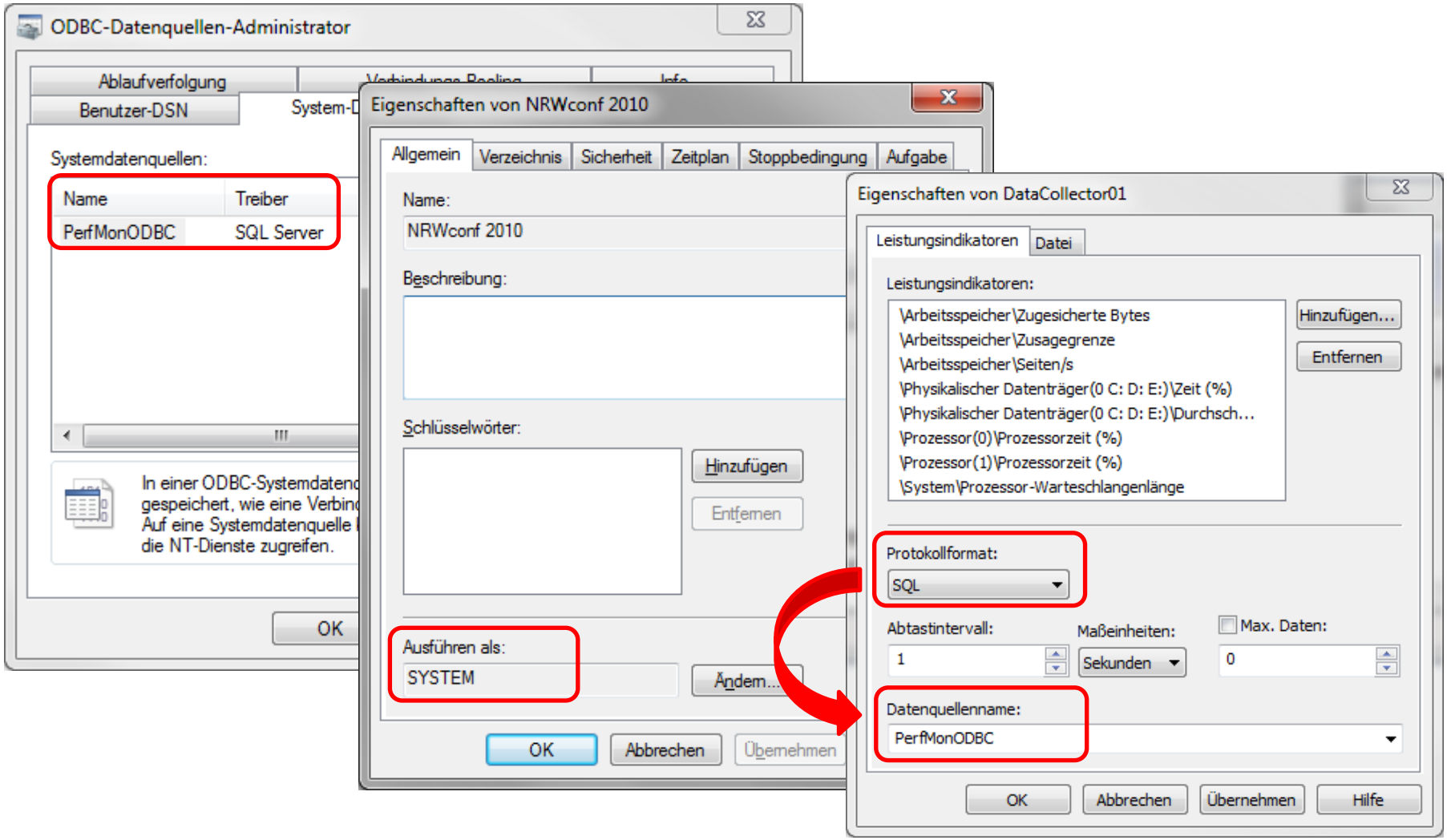
Sammeln von Performance-Daten

- PerfMon + CSV = ein weiterer Arbeitsschritt
  - Zusammenführen aller CSV-Dateien
  - Konvertieren von Zeitstempeln

- Für Faule: PerfMon + DB
  - Gemeinsame Datenbasis
  - Einfache Analyse mit SQL
  - Detaillierte Beschreibung: mein Vortrag im letzten Jahr

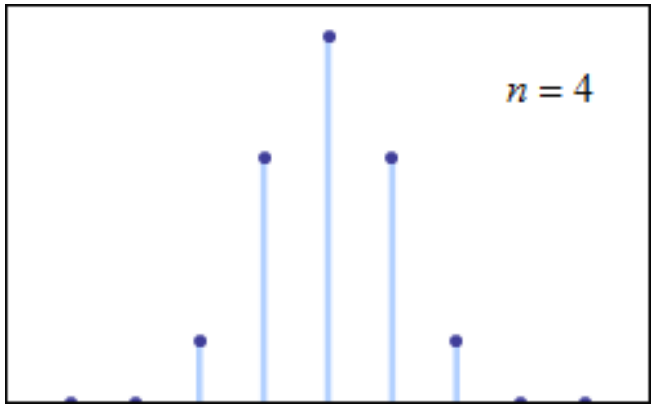
  - Demo
    - Datenbankverbindung
    - Datenbankschema
    - Sichten (PerfMon 1-3)

PerfMon mit Datenbank



Grundlagen der Statistik

- Normalverteilung
  - Der Name ist Programm!
  - Normalerweise sind Stichproben so verteilt
  - Grundlage für viele statistische Analysen



- Glockenform
  - Achsensymmetrisch
  - Höhe und Stauchung hängen von den Daten ab
- Mehr Stichproben ➔ bessere Annäherung der Glockenform
  - Viele Stichproben ermöglichen valide Analyse
- Demo: Histogramm meiner Zeiterfassung

Mittelwerte

- Arithmetisches Mittel
  - Summe geteilt durch Anzahl
  - Gern genommenes Mittelmaß
  - Excel: MITTELWERT()
- Median
  - Mittlerer Wert der sortierten Stichproben
  - Excel: MEAN()
- Besonderheit
  - In normalverteilten Stichproben sind die beiden Mittelwerte gleich
  - Bei Euch nicht? ➔ Mehr Stichproben!

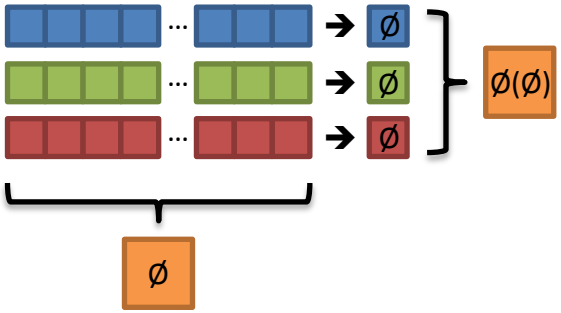
Der Mittelwert – die Mimose

- Ausreißer verfälschen den Mittelwert
  - Wenige Werte deutlich außerhalb der Glocke
  - Verzerrung der Glocke
- Demo
  - Verfälschung des Mittelwerts mit Ausreißern (Outliers 1-2)
- Abhilfe: Ausschließen von Ausreißern
  - SQL: Filtern
  - Excel: MITTELWERTWENN()
- Umso größer der Unterschied zwischen Mittelwert und Median, desto mehr Ausreißer verfälschen die Glocke

Reduktion der Datenmenge

- Aggregation: Speichern von Mittelwerten
  - Messintervall 5 Sekunden
  - Aggregation zu stündlichen Mittelwerten
  - Reduktion der Datenmenge auf 20%

- Arbeiten mit Mittelwerten von Mittelwerten
  - Im Allgemeinen nicht korrekt
  - Mathematisch nur korrekt,
    wenn alle Datensätze gleich groß sind
  - Ø(Ø(An), Ø(Am)) = Ø(An, Am) ➔ n=m



- Demo
  - Auswertung von zwei ungleich großen Datensätzen (Average.sql)

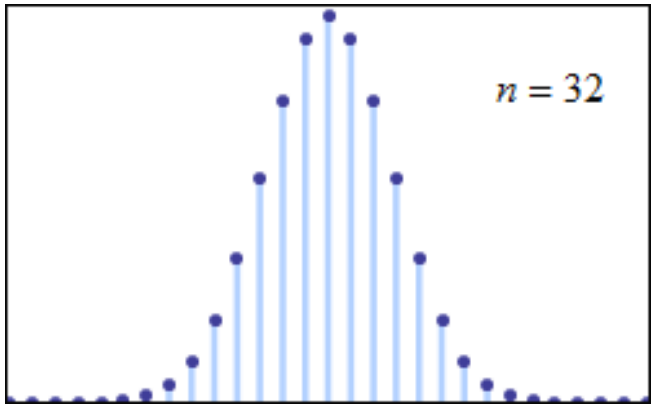
Standardabweichung

- Wie sehr streuen die Stichproben um den Mittelwert?
  - 68,27% aller Werte liegen innerhalb der Standardabweichung
  - 95,45% aller Werte liegen innerhalb von 2x STABW
  - 99,73% aller Werte liegen innerhalb von 3x STABW
  - Offensichtlich: Kleine STABW sind besser

- Excel kennt STABW.N() und STABW.S()
  - STABW.S ist für eine kleine Anzahl von Stichproben
  - STABW.N ist für eine große Anzahl von Stichproben
  - Fehler von STABW.S: $\frac{n}{n-1}$
    - Konverviert gegen 1 für große n: $\lim_{n\to\infty} \frac{n}{n-1} = 1$

Histogramme

- Zweck
  - Erkennung der Häufigkeit von Werten
  - Identifizierung von Ausreißern



- Demo
  - Histogramme mit Excel

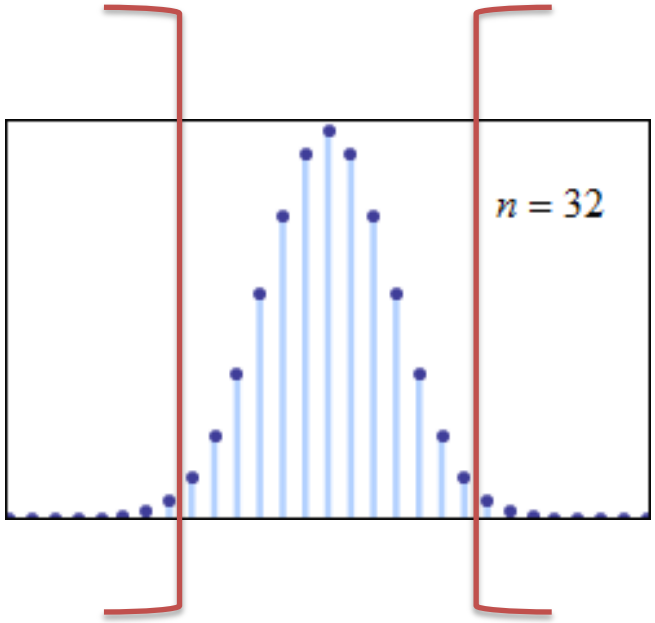
- Demo
  - Histogramme mit SQL (Histogram 1-2)

Hausputz im Datensatz

- Systemimmanente Fehler vs. Ausreißer
  - Manchmal liefert PerfMon unmögliche Werte
  - Beispielsweise: Latenz kleiner Null
- Entfernen von systemimmanenten Fehlern
  - Stichhaltige Argumentationskette
  - Filtern
- Entfernen von echten Ausreißer
  - Vorsicht ist geboten
  - Es handelt sich um reelle Werte …
  - … verursacht durch reelle Situationen
  - Ursachenforschung trotz Bereinigung des Datensatzes

Gestutztes Mittel

- Entfernen der größten und kleinsten Werte
  - Dieselbe Anzahl oder derselbe Anteil
- Achtung: Eventuell werden reale Werte verworfen
  - Analyse erforderlich
- Demo
  - Stutzen mit SQL (TruncMean 1-5)
- Demo
  - Das gestutzte Mittel als Sicht auf Tabellen (TruncMean 6-7)



Fazit

- Glocken sind gut
  - Die Normalverteilung ermöglich statistische Analysen
  - Mittelwert und Median liegen nah beieinander
- Mittelwerte sind Mimosen
  - Ausreißer verfälschen den Mittelwert
- Histogramme sind auch gut
  - Ausreißer werden enttarnt
- Gestutztes Mittel einsetzen
  - Statistisch korrekte Bereinigung

Visualisieren

- Excel
  - Nicht importieren sondern verknüpfen
- SQL Server
  Business Intelligence Development Studio (BIDS)
  - Basiert auf Visual Studio
  - Vorlagen für Reports
- SQL Server Reporting Services
  Report Builder
  - Standalone-Werkzeug zur Report-Entwicklung
  - Aufruf von http://<SERVERNAME>/Reports

NRW Conf
Moving the IT Community