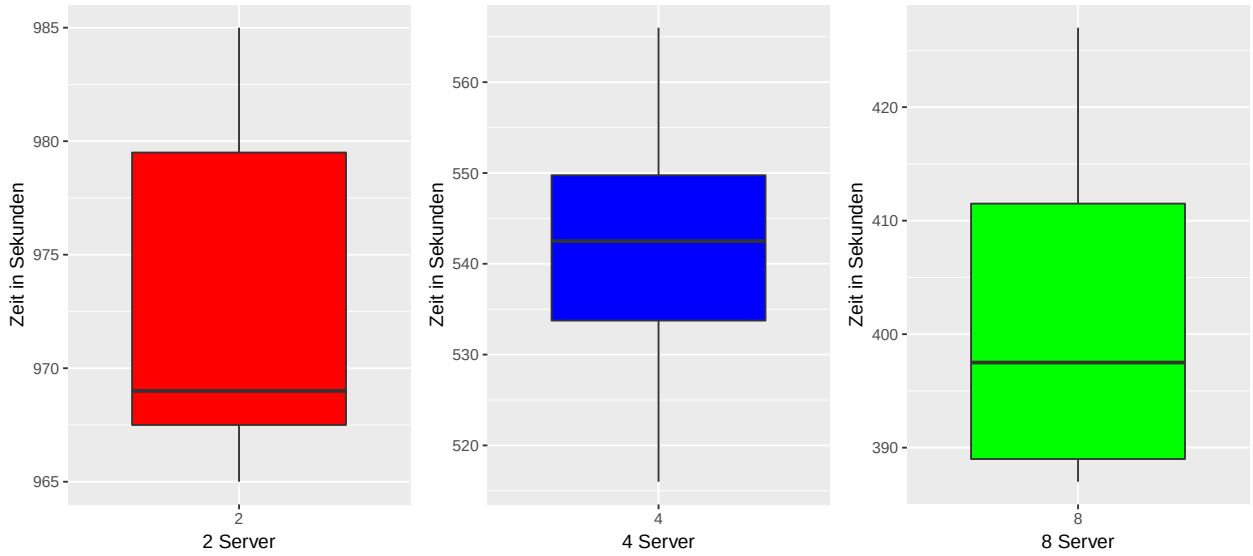
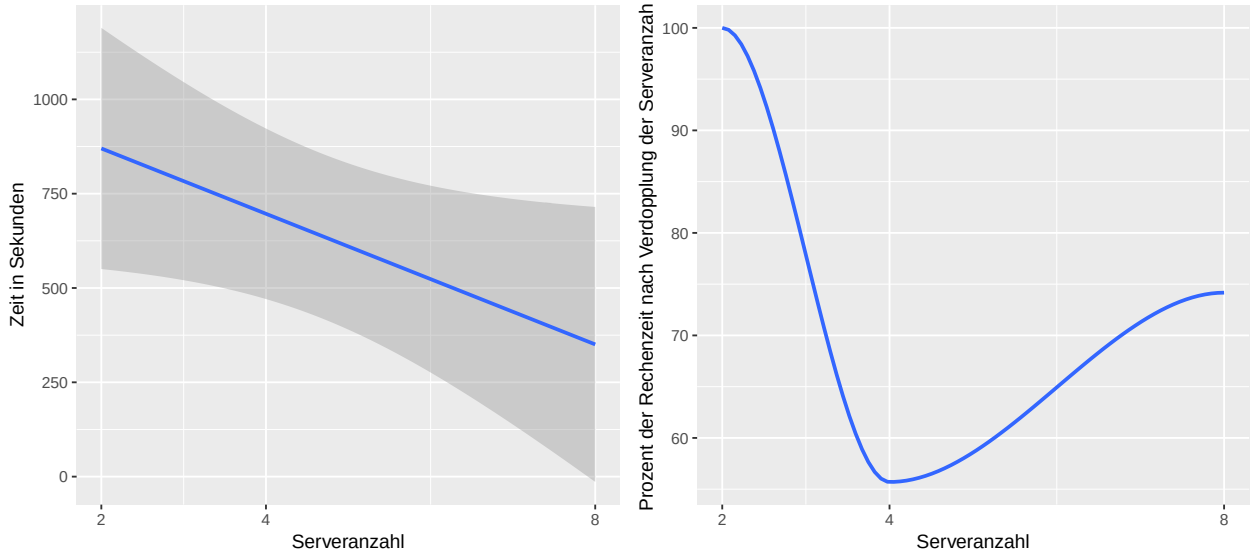
Praktikum 5

Wir haben 40GB zufälligen Text erzeugt und anschließend die Anzahl an Wörtern mit Hadoop mittels WordCount gezählt und die Zeiten von Anfang der Bearbeitung der Abfrage (MapReduce Start) bis Ende der Bearbeitung und erhalt des Ergebnisses (MapReduce Ende) gemessen und gespeichert. Die Zeiten haben wir in Sekunden gemessen und folgende Ergebnisse erhalten.



In den Boxplots können wir die genaue Verteilung der gemessenen Zeiten in Abhängigkeit der Serveranzahl sehen. Wie wir sehen können führte eine Erhöhung der Serveranzahl zu einer signifikanten Reduktion der Rechenzeit. Dies analysieren wir genauer.



In der linken Grafik sehen wir die Rechenzeit in Abhängigkeit der Serveranzahl. Wie wir sehen können gibt es einen signifikaten sinkenden Trend. In der rechten Grafik sehen wir die Prozentuale Reduktion der Serverzeit nach jeweiliger Verdopplung der Serveranzahl. Wie wir sehen können Bedarf es genauer Planung, wie viel Zeitgewinn und was für ein Kostenleistungsverhältnis man erreichen möchte. Bei Verdopplung von 2 auf 4 Server hat sich die Rechenzeit fast halbiert, während die Rechenzeit nach Verdopplung von 4 auf 8 nur um etwa 25% gesenkt wurde. Je nach Anwendung wäre es also sinnvoll noch zu messen ab welcher Serveranzahl die Performancesteigerung nicht mehr mit den Kosten rechtfertigbar ist.

Alles in allem können wir somit belegen, dass die Rechenzeit definitiv mit Steigender Serveranzahl und horizontaler Saklierung verbessert werden kann ohne, dass die Leistung der Server selbst gesteigert werden muss. Da wir dynamisch skalieren können bietet uns dies somit den besonderen Vorteil, dass wir solange gleiche Maschinen dazu kaufen können, bis die gewünschte Performance erreicht hat. Da die Maschine als Klone erstellt werden können sparen wir sowohl Konfigurationsaufwand als auch deutliche Einsparung bei den Maschinenkosten, da doppelte Leitung eines Geräts im Regelfall ein Vielfaches mehr kostet während bei mehreren neuen Maschinen die Kosten linear skalieren und wir eventuell noch Mengenrabatt aushandeln können.

Besonders durch die Skalierbarkeit und den einfachen und Austauschbaren Performancegewinn durch horizontale Erweiterung sprechen wir eine klare Empfehlung für eben diese mittels Hadoop aus.