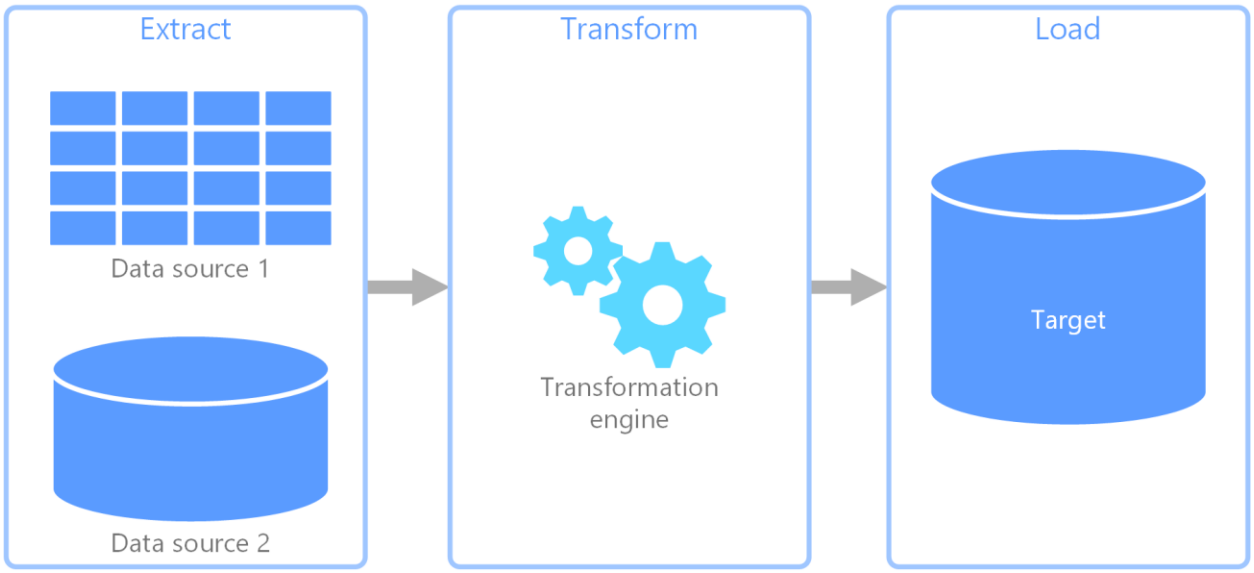
ETL PROJECT

Group – 7 (James Akerman, Gianna Abono, Lino Zhang, Joseph Tatapudi)



Introduction

We are a financial advisory firm. One of our major clients is interested in purchasing shares from either the **Commonwealth Bank of Australia (CBA)** or the **Australia and New Zealand Banking Group (ANZ)**, and wants our advice on which stock to choose.

In order to conduct the analysis that we need to do to properly advise them on which one to purchase, we have designed and created an ETL process to collect all the most recent CBA and ANZ financial data from the Yahoo Finance website.

This process includes web scraping, data transformation, and database creating and loading.

By clicking on 'Run' in the **Complete ETL Process** notebook, the complete ETL process automatically for both the CBA and ANZ stocks. The ETL processes runs for

each of the collections in the data for both stocks. Once the process is complete, the fresh data will be available in the database, which can be queried and analyzed in order for us to provided our informed recommendation our client regarding which stock they should purchase.

Justification for the data chosen:

To be able to analyze the value of the stocks and give our informed recommendation about which stock to buy, we needed the following information:

- **Summary Data:** gives us an overall snapshot of all the stocks. PE ratio helps us determine the Value of a Stock. Market capitalization gives us the size of a company. This information can be used by us to assess which of these stocks is better value for money.
- **Income Statement:** shows us Net income and Revenue which helps us determine financial strength of both companies.
- **Stock History**: shows information about the changes in each company's share price overtime such as price changes, current trading price, historical highs and lows, which allows us to assess which company's shares have performed better over time.
- **Balance Sheet:** gives us information about each company's assets and liabilities which helps us to assess which company has more assets or more debt and thus which one might be a better buy.
- **Cash Flow:** helps us to analyze financial health of each company by allowing us to compare their sources of income and expenses.

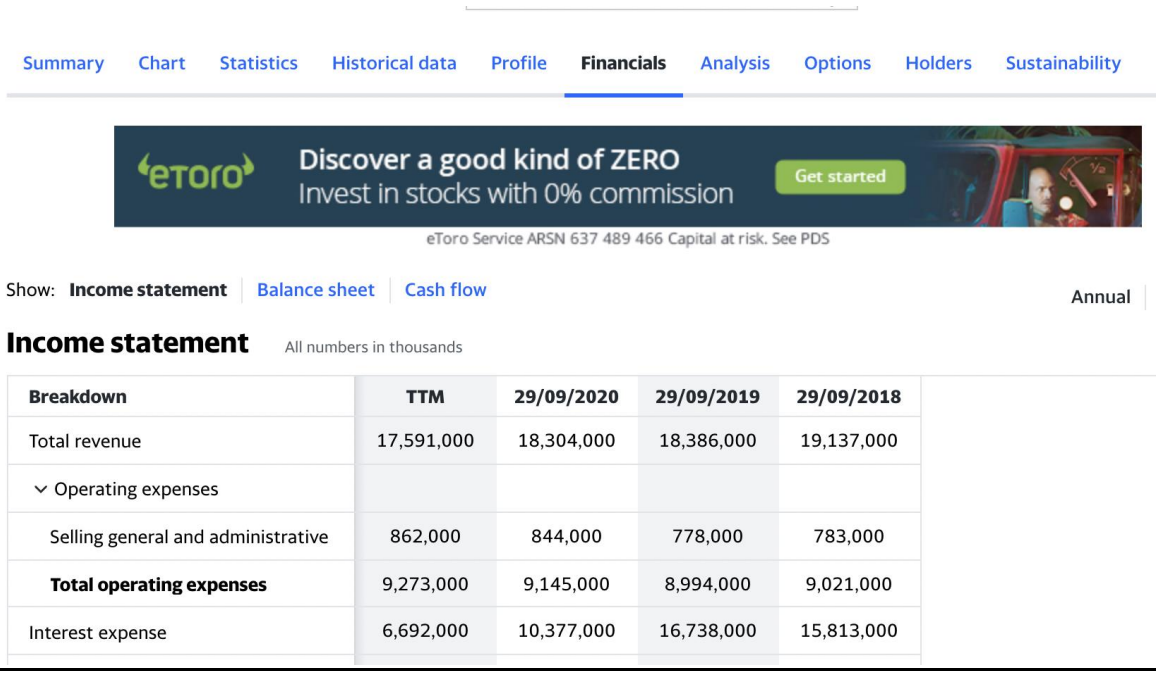
Extracting data

Data Resource:

Web Scraping from the links below:

https://au.finance.yahoo.com/quote/CBA.AX/financials?p=CBA.AX

https://au.finance.yahoo.com/quote/ANZ.AX/financials?p=ANZ.AX

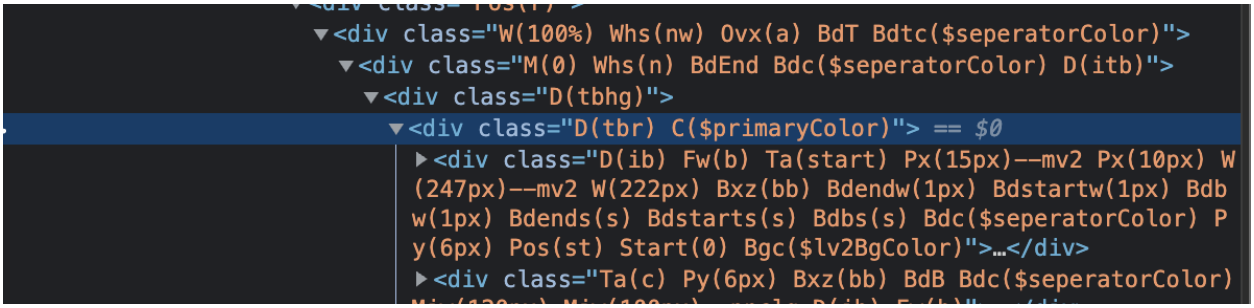


Breakdown	TTM	29/09/2020	29/09/2019	29/09/2018
Total revenue	17,591,000	18,304,000	18,386,000	19,137,000
Operating expenses				
Selling general and administrative	862,000	844,000	778,000	783,000
Total operating expenses	9,273,000	9,145,000	8,994,000	9,021,000
Interest expense	6,692,000	10,377,000	16,738,000	15,813,000

Under Financials, there are three sheets: Income Statement, Balance sheet, and Cash flow.

We applied three methods to scrape 'Fresh' data from the website

1. We wrote scraping scripts using Beautiful Soup to scrape tables (Extract)
2. We also scraped by selecting targeted 'class' to get information (Extract)
3. Finally, we scraped the csv file then saved it to the local directory by running the extract process. (Extra → Load)



Transforming data

Data Cleaning:

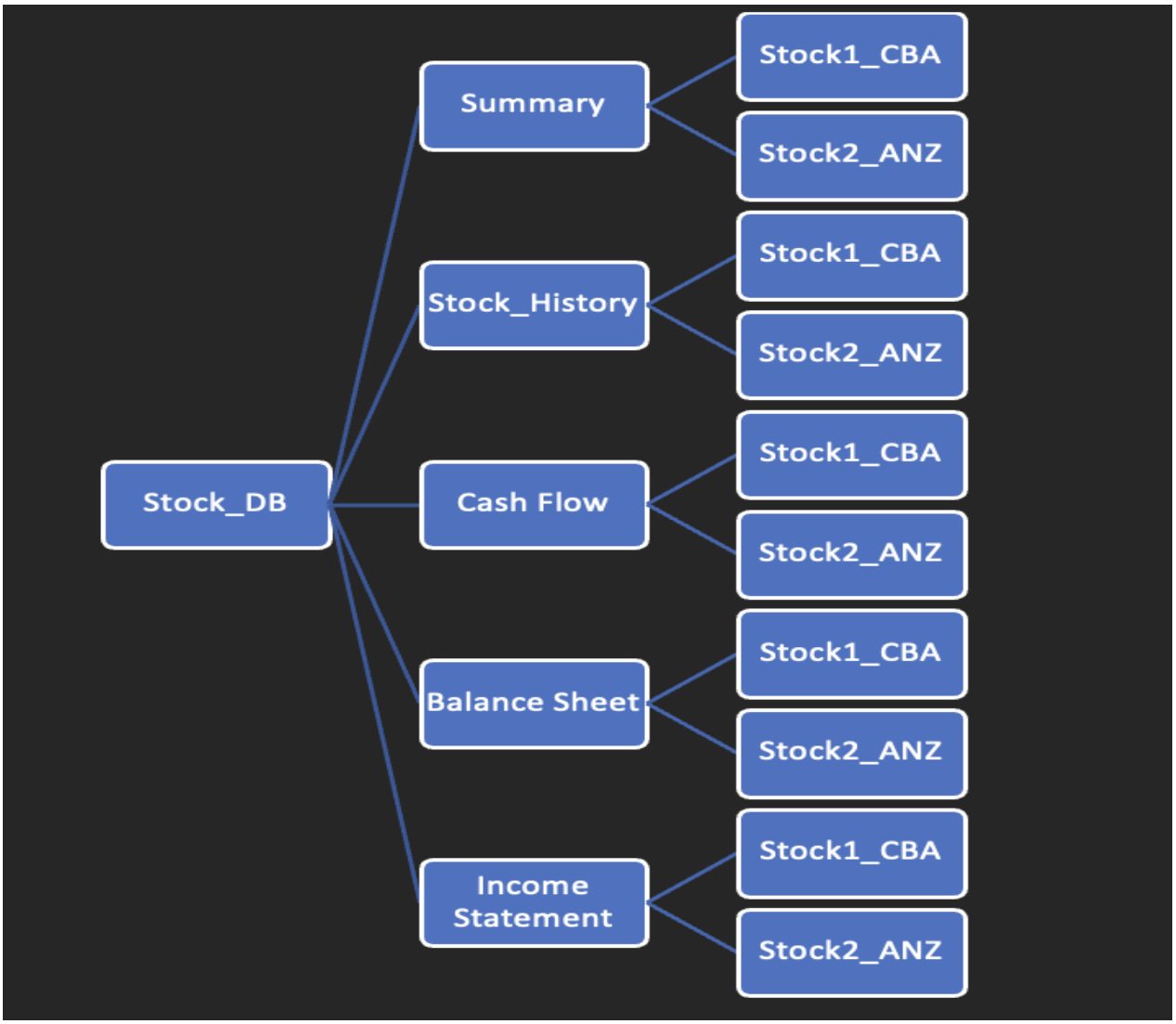
We cleaned raw data before loading them into the database. Methods that we applied in cleaning process include: removing na/null values, renaming and filtering table columns; formatting dates strings in year-month-day; replacing empty values with 0s; round up decimal places; and setting indexes. We sorted cleaned data into the dictionaries with desired hierarchy before we load them in the database.

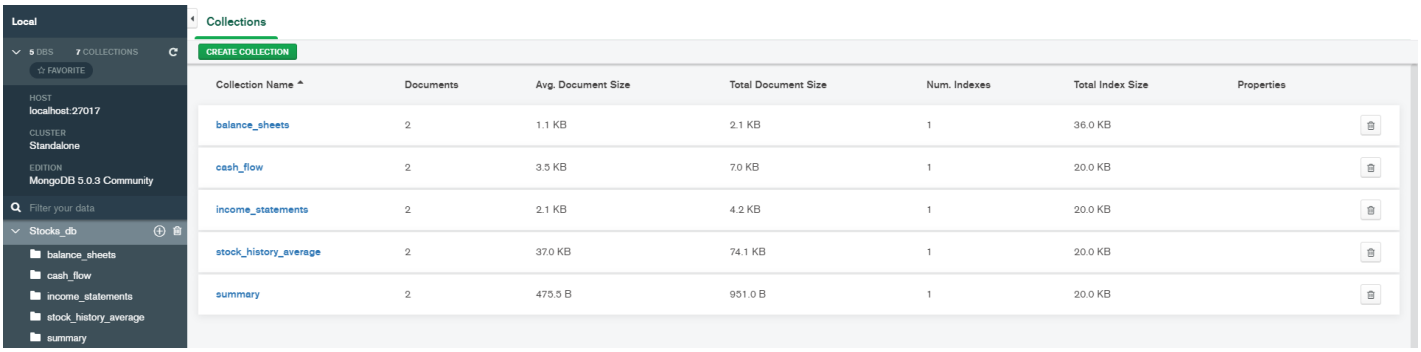
Loading data

We created the connection between the python file and database. We transferred our final data output to **Stocks_DB** by creating a new database in python script.

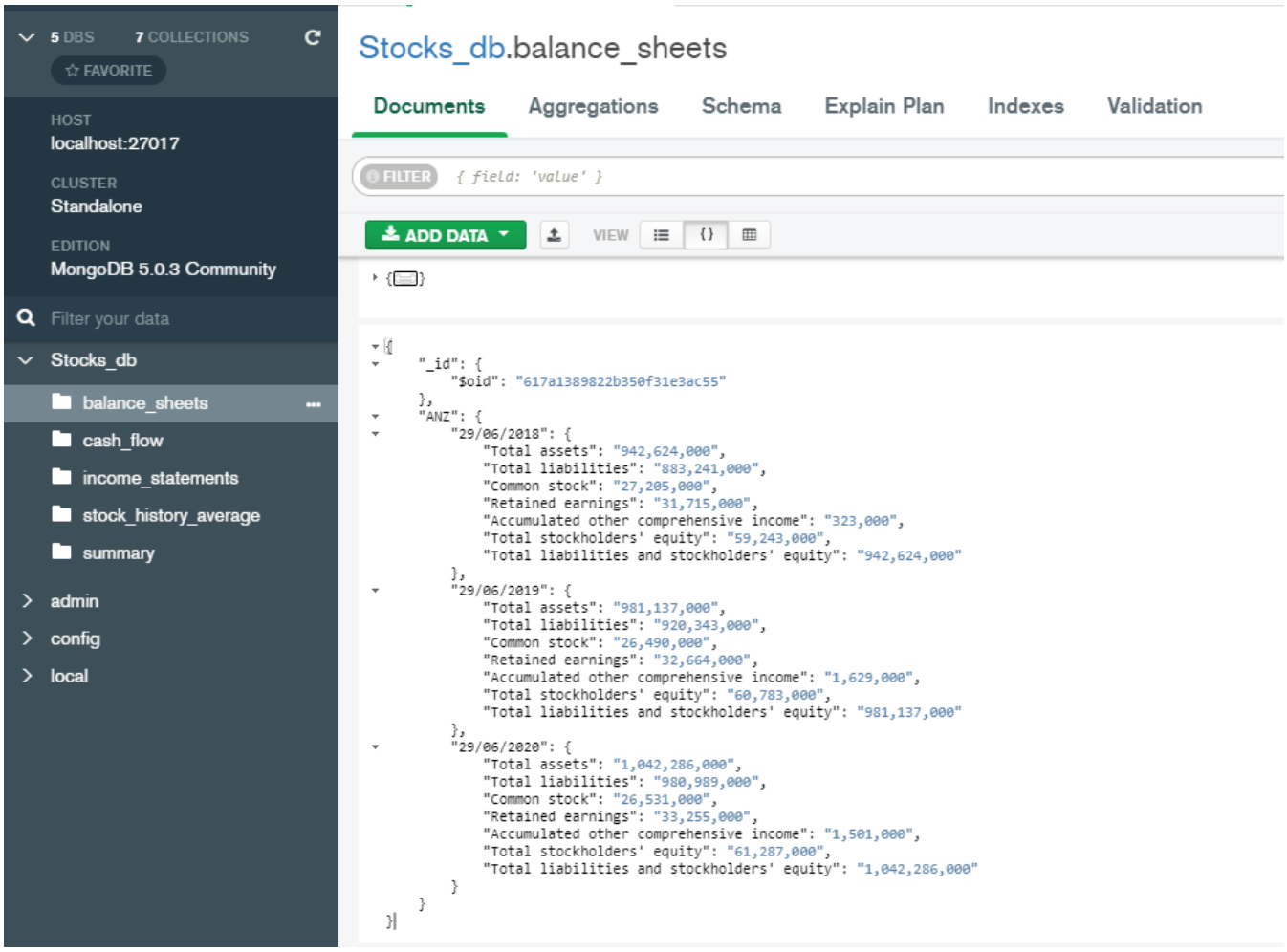
Below are the screenshots of our database and table collections

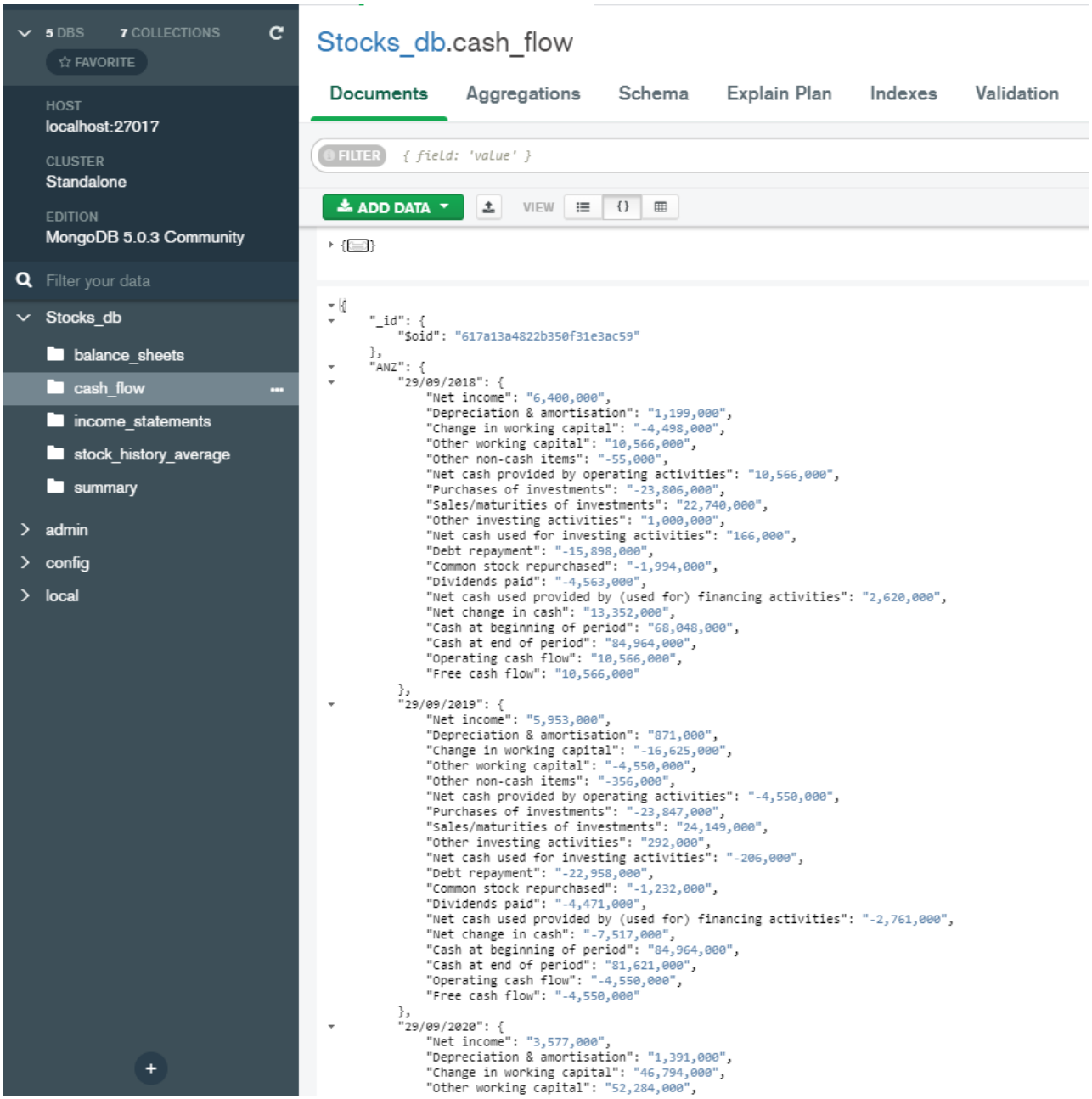
Breakdown structure:

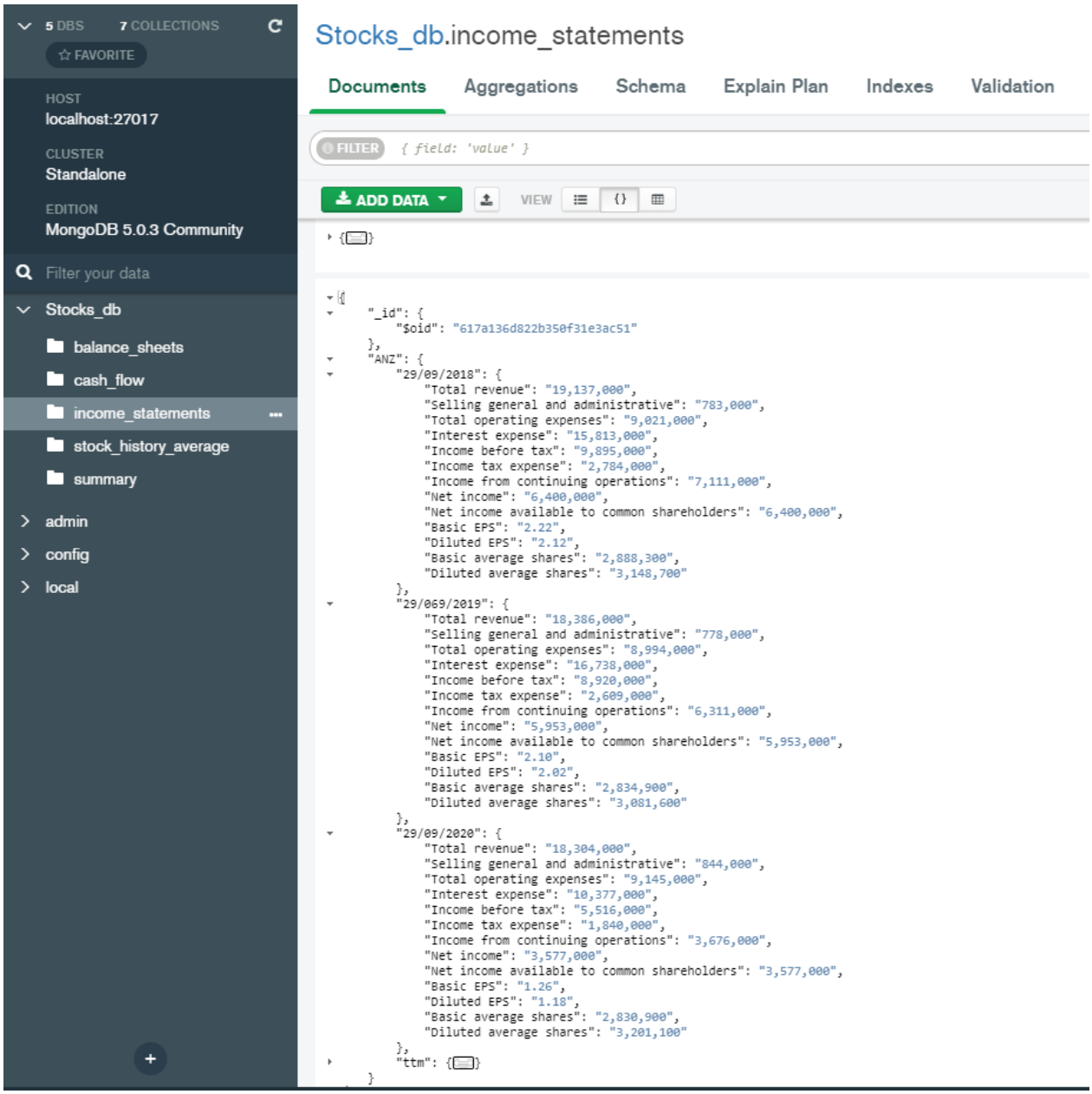


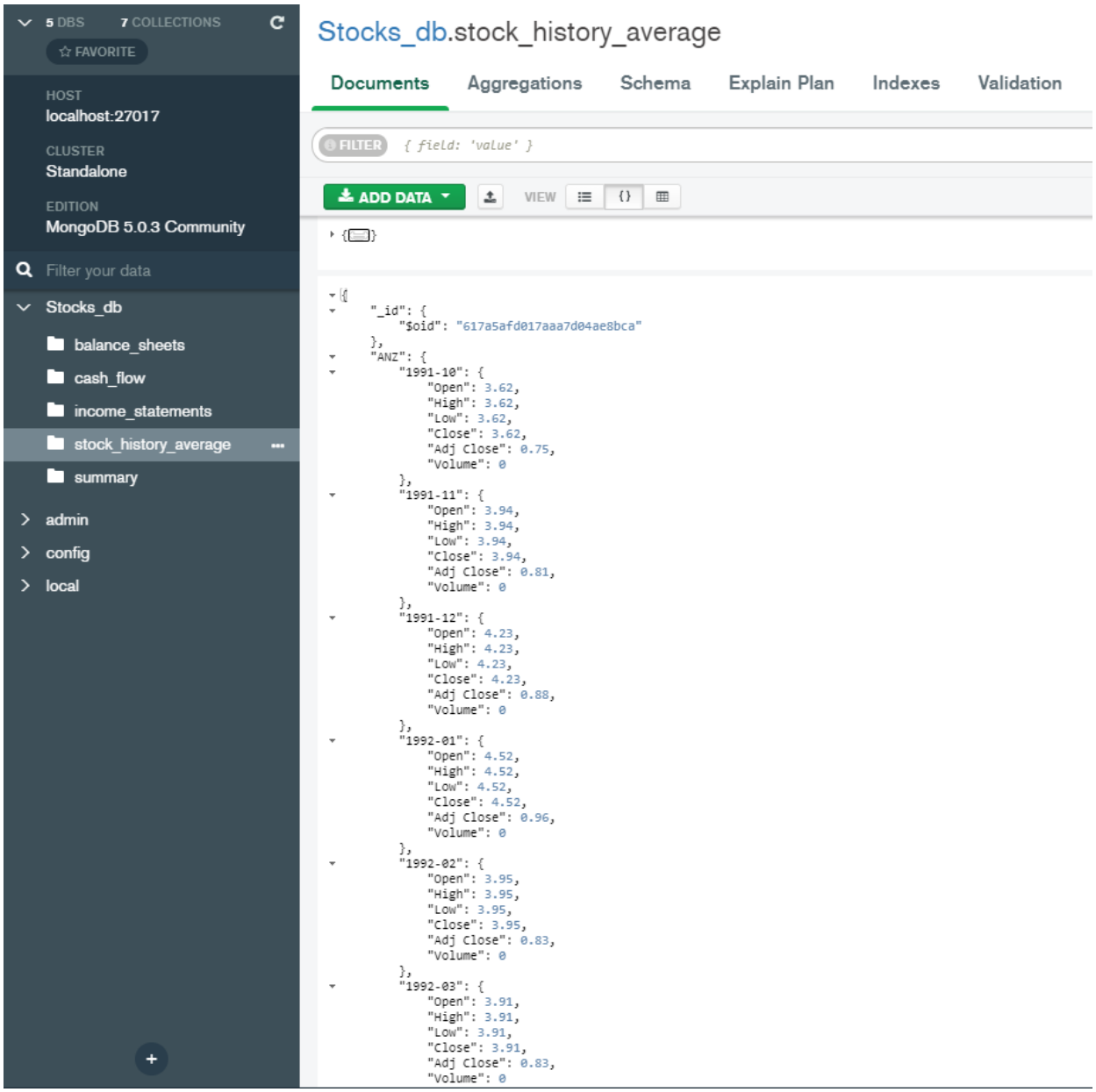


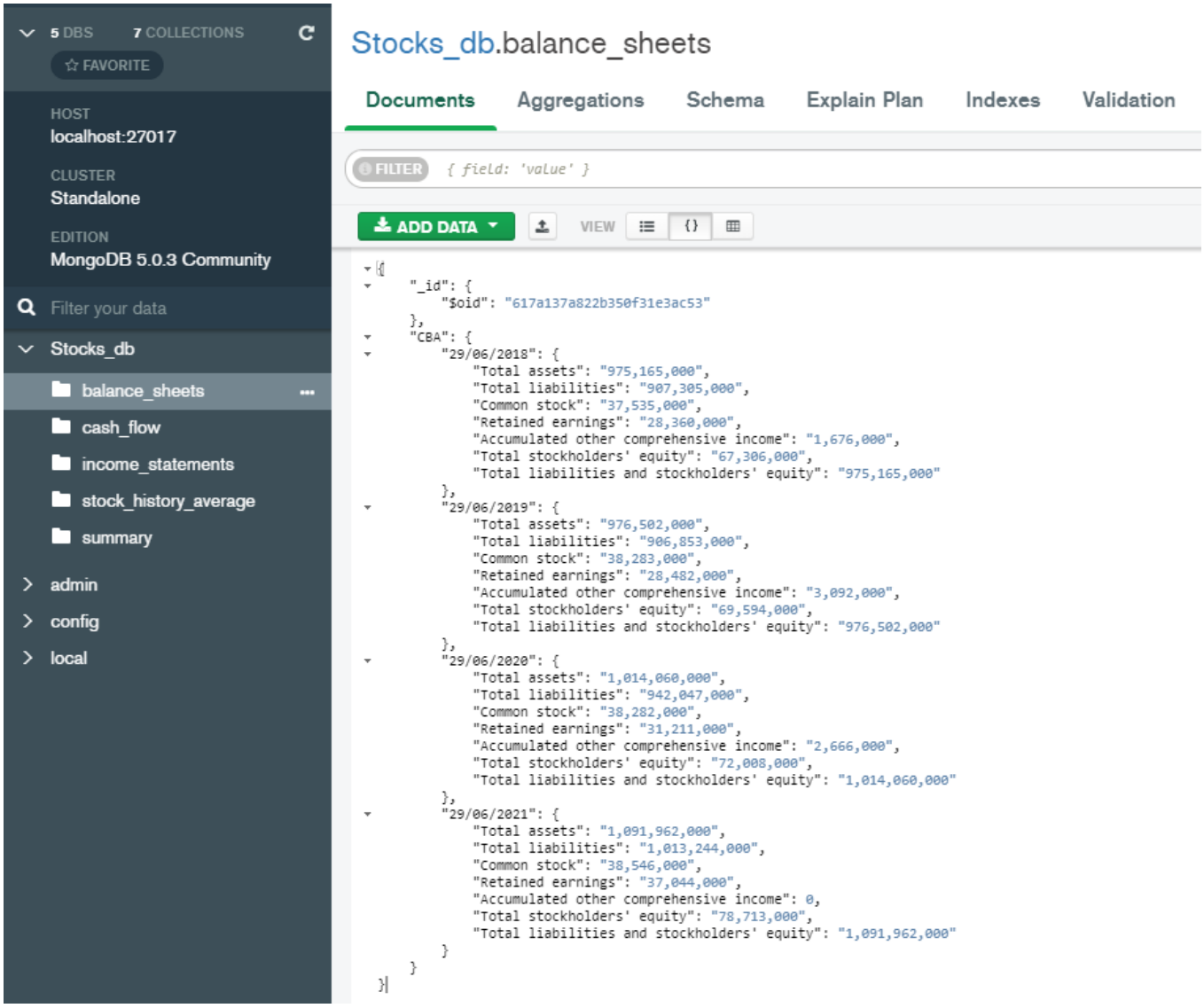
Collection Name	Documents	Avg. Document Size	Total Document Size	Num. Indexes	Total Index Size	Properties
balance_sheets	2	1.1 KB	2.1 KB	1	36.0 KB	
cash_flow	2	3.5 KB	7.0 KB	1	20.0 KB	
income_statements	2	2.1 KB	4.2 KB	1	20.0 KB	
stock_history_average	2	37.0 KB	74.1 KB	1	20.0 KB	
summary	2	475.5 B	951.0 B	1	20.0 KB	

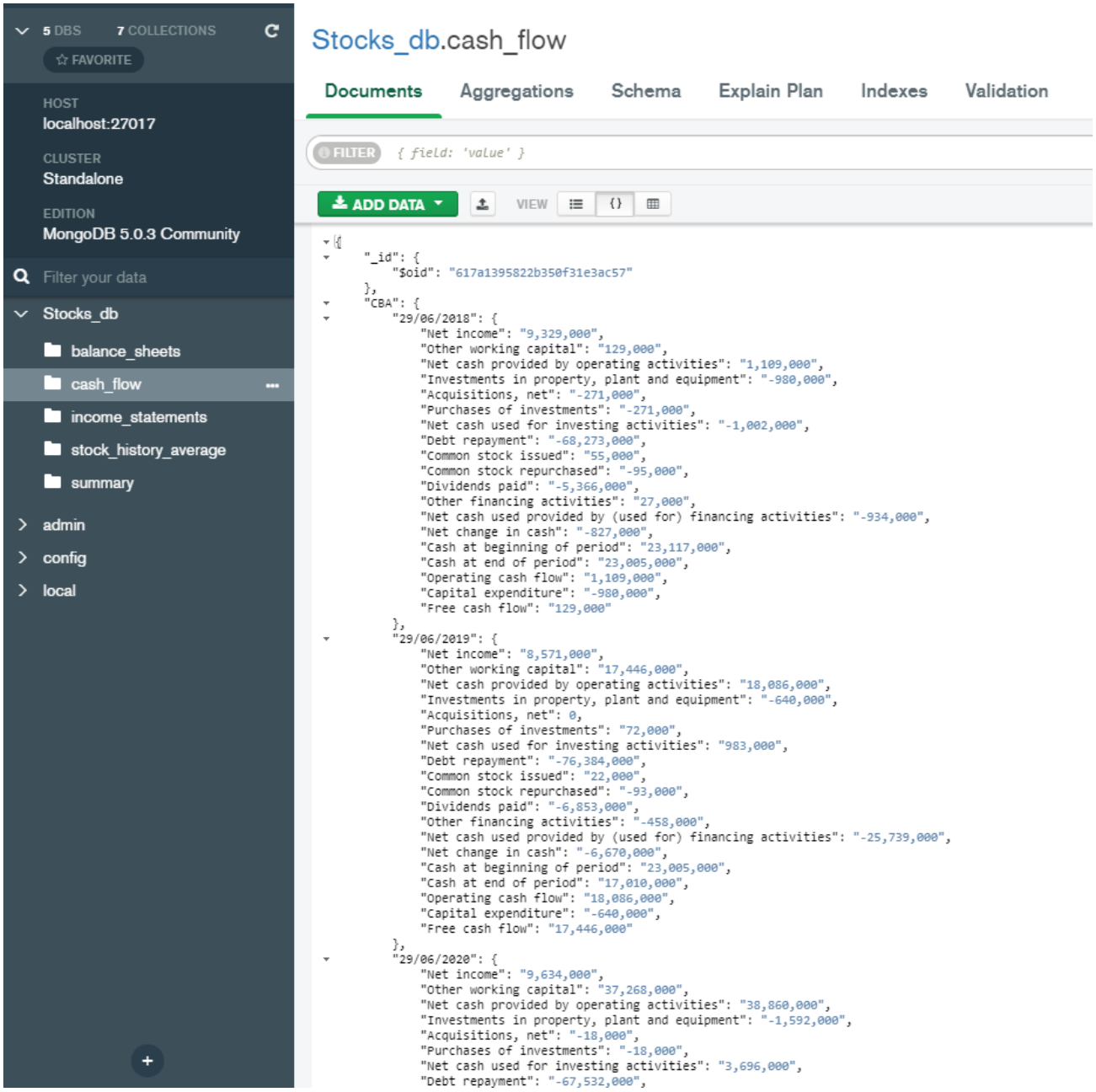

5 DBS
7 COLLECTIONS
FAVORITE
HOST
localhost:27017
CLUSTER
Standalone
EDITION
MongoDB 5.0.3 Community
Filter your data
Stocks_db
balance_sheets
cash_flow
income_statements
stock_history_average
summary
admin
config
local
Stocks_db.balance_sheets
Documents
Aggregations
Schema
Explain Plan
Indexes
Validation
FILTER
{ field: 'value' }
ADD DATA
VIEW
{ "_id": { "$oid": "617a1389822b350f31e3ac55" }, "ANZ": { "29/06/2018": { "Total assets": "942,624,000", "Total liabilities": "883,241,000", "Common stock": "27,205,000", "Retained earnings": "31,715,000", "Accumulated other comprehensive income": "323,000", "Total stockholders' equity": "59,243,000", "Total liabilities and stockholders' equity": "942,624,000" }, "29/06/2019": { "Total assets": "981,137,000", "Total liabilities": "920,343,000", "Common stock": "26,490,000", "Retained earnings": "32,664,000", "Accumulated other comprehensive income": "1,629,000", "Total stockholders' equity": "60,783,000", "Total liabilities and stockholders' equity": "981,137,000" }, "29/06/2020": { "Total assets": "1,042,286,000", "Total liabilities": "980,989,000", "Common stock": "26,531,000", "Retained earnings": "33,255,000", "Accumulated other comprehensive income": "1,501,000", "Total stockholders' equity": "61,287,000", "Total liabilities and stockholders' equity": "1,042,286,000" } } }

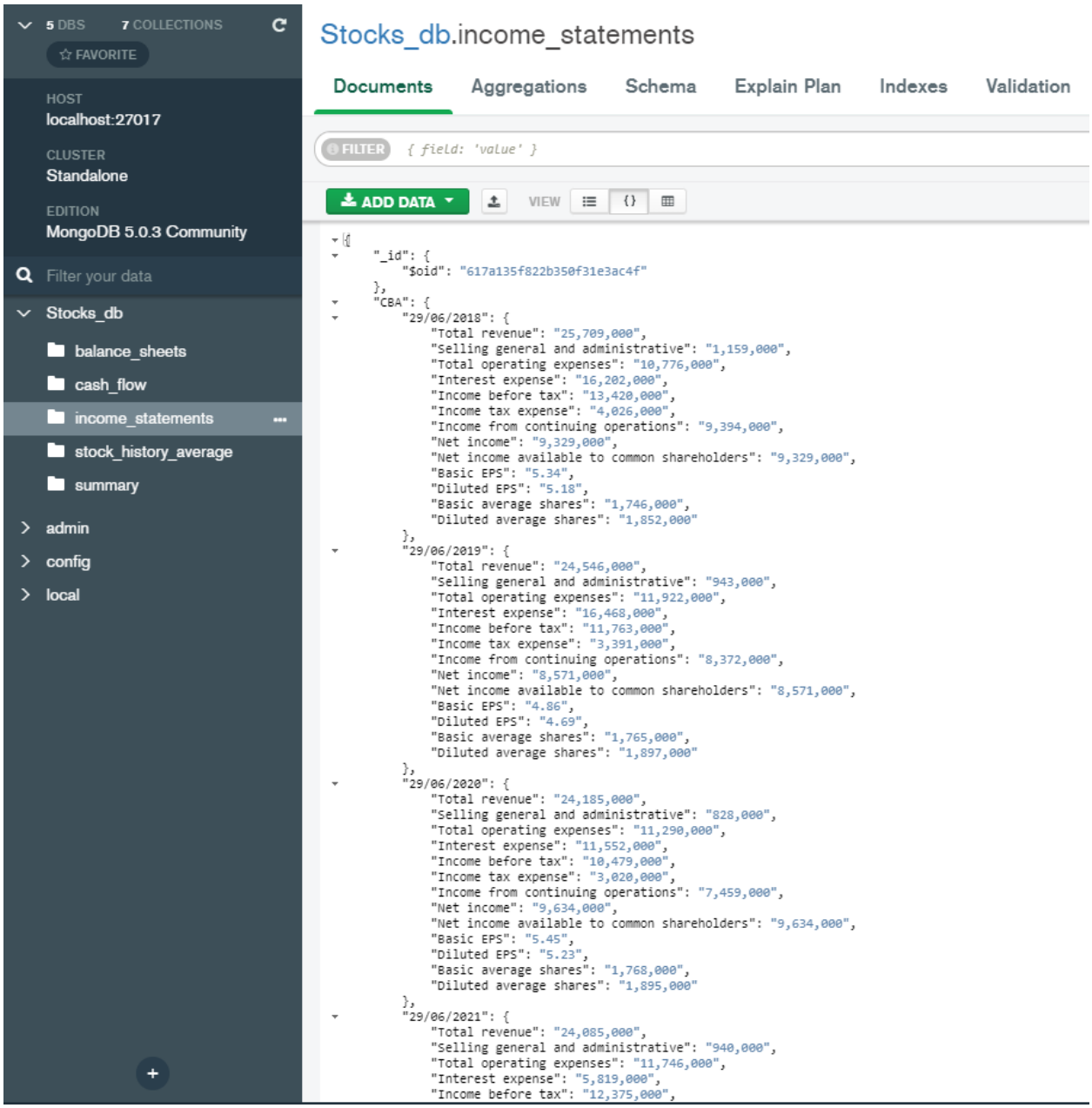

5 DBS
7 COLLECTIONS
FAVORITE
HOST
localhost:27017
CLUSTER
Standalone
EDITION
MongoDB 5.0.3 Community
Filter your data
Stocks_db
balance_sheets
cash_flow
income_statements
stock_history_average
summary
admin
config
local
Stocks_db.cash_flow
Documents
Aggregations
Schema
Explain Plan
Indexes
Validation
FILTER
{ field: 'value' }
ADD DATA
VIEW
{}
{
"_id": {
"$oid": "617a13a4822b350f31e3ac59"
},
"ANZ": {
"29/09/2018": {
"Net income": "6,400,000",
"Depreciation & amortisation": "1,199,000",
"Change in working capital": "-4,498,000",
"Other working capital": "10,566,000",
"Other non-cash items": "-55,000",
"Net cash provided by operating activities": "10,566,000",
"Purchases of investments": "-23,806,000",
"Sales/maturities of investments": "22,740,000",
"Other investing activities": "1,000,000",
"Net cash used for investing activities": "166,000",
"Debt repayment": "-15,898,000",
"Common stock repurchased": "-1,994,000",
"Dividends paid": "-4,563,000",
"Net cash used provided by (used for) financing activities": "2,620,000",
"Net change in cash": "13,352,000",
"Cash at beginning of period": "68,048,000",
"Cash at end of period": "84,964,000",
"Operating cash flow": "10,566,000",
"Free cash flow": "10,566,000"
},
"29/09/2019": {
"Net income": "5,953,000",
"Depreciation & amortisation": "871,000",
"Change in working capital": "-16,625,000",
"Other working capital": "-4,550,000",
"Other non-cash items": "-356,000",
"Net cash provided by operating activities": "-4,550,000",
"Purchases of investments": "-23,847,000",
"Sales/maturities of investments": "24,149,000",
"Other investing activities": "292,000",
"Net cash used for investing activities": "-206,000",
"Debt repayment": "-22,958,000",
"Common stock repurchased": "-1,232,000",
"Dividends paid": "-4,471,000",
"Net cash used provided by (used for) financing activities": "-2,761,000",
"Net change in cash": "-7,517,000",
"Cash at beginning of period": "84,964,000",
"Cash at end of period": "81,621,000",
"Operating cash flow": "-4,550,000",
"Free cash flow": "-4,550,000"
},
"29/09/2020": {
"Net income": "3,577,000",
"Depreciation & amortisation": "1,391,000",
"Change in working capital": "46,794,000",
"Other working capital": "52,284,000",

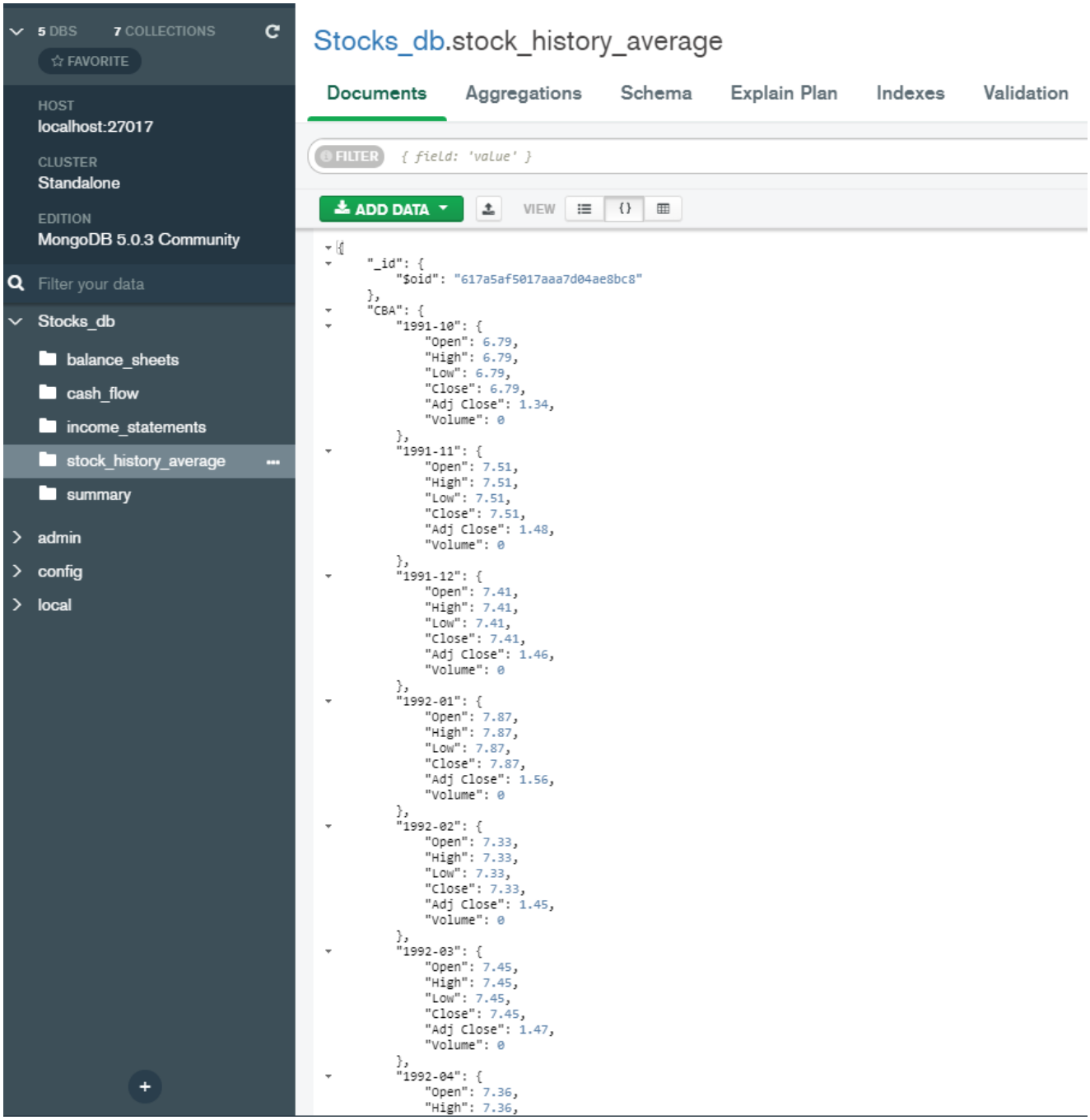

5 DBS
7 COLLECTIONS
FAVORITE
HOST
localhost:27017
CLUSTER
Standalone
EDITION
MongoDB 5.0.3 Community
Filter your data
Stocks_db
balance_sheets
cash_flow
income_statements
stock_history_average
summary
admin
config
local
Stocks_db.income_statements
Documents
Aggregations
Schema
Explain Plan
Indexes
Validation
FILTER
{ field: 'value' }
ADD DATA
VIEW
{}
{
"_id": {
"$oid": "617a136d822b350f31e3ac51"
},
"ANZ": {
"29/09/2018": {
"Total revenue": "19,137,000",
"Selling general and administrative": "783,000",
"Total operating expenses": "9,021,000",
"Interest expense": "15,813,000",
"Income before tax": "9,895,000",
"Income tax expense": "2,784,000",
"Income from continuing operations": "7,111,000",
"Net income": "6,400,000",
"Net income available to common shareholders": "6,400,000",
"Basic EPS": "2.22",
"Diluted EPS": "2.12",
"Basic average shares": "2,888,300",
"Diluted average shares": "3,148,700"
},
"29/069/2019": {
"Total revenue": "18,386,000",
"Selling general and administrative": "778,000",
"Total operating expenses": "8,994,000",
"Interest expense": "16,738,000",
"Income before tax": "8,920,000",
"Income tax expense": "2,609,000",
"Income from continuing operations": "6,311,000",
"Net income": "5,953,000",
"Net income available to common shareholders": "5,953,000",
"Basic EPS": "2.10",
"Diluted EPS": "2.02",
"Basic average shares": "2,834,900",
"Diluted average shares": "3,081,600"
},
"29/09/2020": {
"Total revenue": "18,304,000",
"Selling general and administrative": "844,000",
"Total operating expenses": "9,145,000",
"Interest expense": "10,377,000",
"Income before tax": "5,516,000",
"Income tax expense": "1,840,000",
"Income from continuing operations": "3,676,000",
"Net income": "3,577,000",
"Net income available to common shareholders": "3,577,000",
"Basic EPS": "1.26",
"Diluted EPS": "1.18",
"Basic average shares": "2,830,900",
"Diluted average shares": "3,201,100"
},
"ttm": {}
}

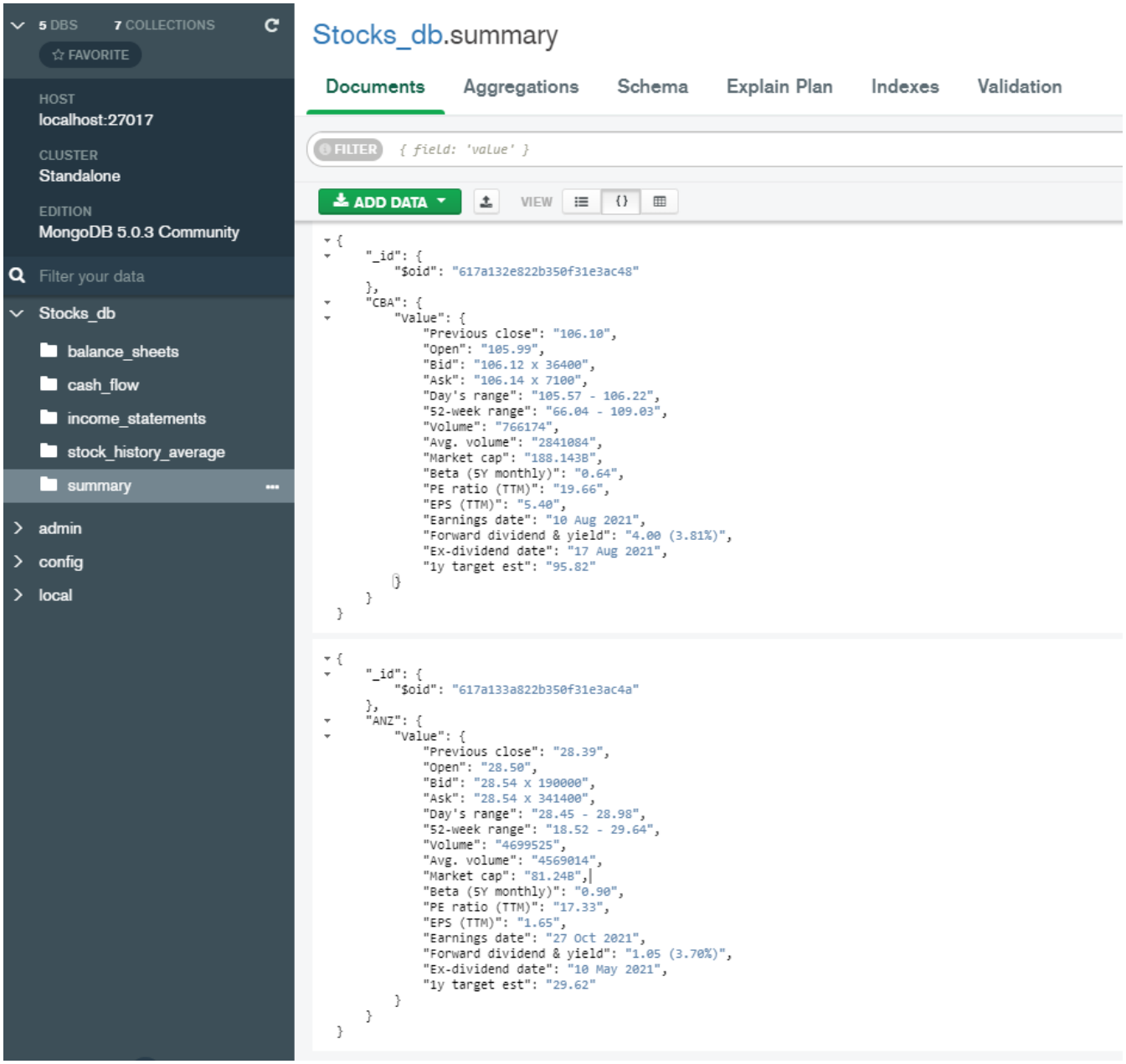

5 DBS
7 COLLECTIONS
FAVORITE
HOST
localhost:27017
CLUSTER
Standalone
EDITION
MongoDB 5.0.3 Community
Filter your data
Stocks_db
balance_sheets
cash_flow
income_statements
stock_history_average
summary
admin
config
local
Stocks_db.stock_history_average
Documents
Aggregations
Schema
Explain Plan
Indexes
Validation
FILTER
{ field: 'value' }
ADD DATA
VIEW
{ "_id": { "$oid": "617a5afd017aaa7d04ae8bca" }, "ANZ": { "1991-10": { "Open": 3.62, "High": 3.62, "Low": 3.62, "Close": 3.62, "Adj Close": 0.75, "Volume": 0 }, "1991-11": { "Open": 3.94, "High": 3.94, "Low": 3.94, "Close": 3.94, "Adj Close": 0.81, "Volume": 0 }, "1991-12": { "Open": 4.23, "High": 4.23, "Low": 4.23, "Close": 4.23, "Adj Close": 0.88, "Volume": 0 }, "1992-01": { "Open": 4.52, "High": 4.52, "Low": 4.52, "Close": 4.52, "Adj Close": 0.96, "Volume": 0 }, "1992-02": { "Open": 3.95, "High": 3.95, "Low": 3.95, "Close": 3.95, "Adj Close": 0.83, "Volume": 0 }, "1992-03": { "Open": 3.91, "High": 3.91, "Low": 3.91, "Close": 3.91, "Adj Close": 0.83, "Volume": 0


5 DBS
7 COLLECTIONS
FAVORITE
HOST
localhost:27017
CLUSTER
Standalone
EDITION
MongoDB 5.0.3 Community
Filter your data
Stocks_db
balance_sheets
cash_flow
income_statements
stock_history_average
summary
admin
config
local
Stocks_db.balance_sheets
Documents
Aggregations
Schema
Explain Plan
Indexes
Validation
FILTER
{ field: 'value' }
ADD DATA
VIEW
{
"_id": {
"$oid": "617a137a822b350f31e3ac53"
},
"CBA": {
"29/06/2018": {
"Total assets": "975,165,000",
"Total liabilities": "907,305,000",
"Common stock": "37,535,000",
"Retained earnings": "28,360,000",
"Accumulated other comprehensive income": "1,676,000",
"Total stockholders' equity": "67,306,000",
"Total liabilities and stockholders' equity": "975,165,000"
},
"29/06/2019": {
"Total assets": "976,502,000",
"Total liabilities": "906,853,000",
"Common stock": "38,283,000",
"Retained earnings": "28,482,000",
"Accumulated other comprehensive income": "3,092,000",
"Total stockholders' equity": "69,594,000",
"Total liabilities and stockholders' equity": "976,502,000"
},
"29/06/2020": {
"Total assets": "1,014,060,000",
"Total liabilities": "942,047,000",
"Common stock": "38,282,000",
"Retained earnings": "31,211,000",
"Accumulated other comprehensive income": "2,666,000",
"Total stockholders' equity": "72,008,000",
"Total liabilities and stockholders' equity": "1,014,060,000"
},
"29/06/2021": {
"Total assets": "1,091,962,000",
"Total liabilities": "1,013,244,000",
"Common stock": "38,546,000",
"Retained earnings": "37,044,000",
"Accumulated other comprehensive income": 0,
"Total stockholders' equity": "78,713,000",
"Total liabilities and stockholders' equity": "1,091,962,000"
}
}
}


5 DBS
7 COLLECTIONS
FAVORITE
HOST
localhost:27017
CLUSTER
Standalone
EDITION
MongoDB 5.0.3 Community
Filter your data
Stocks_db
balance_sheets
cash_flow
income_statements
stock_history_average
summary
admin
config
local
Stocks_db.cash_flow
Documents
Aggregations
Schema
Explain Plan
Indexes
Validation
FILTER
{ field: 'value' }
ADD DATA
VIEW
{
"_id": {
"$oid": "617a1395822b350f31e3ac57"
},
"CBA": {
"29/06/2018": {
"Net income": "9,329,000",
"Other working capital": "129,000",
"Net cash provided by operating activities": "1,109,000",
"Investments in property, plant and equipment": "-980,000",
"Acquisitions, net": "-271,000",
"Purchases of investments": "-271,000",
"Net cash used for investing activities": "-1,002,000",
"Debt repayment": "-68,273,000",
"Common stock issued": "55,000",
"Common stock repurchased": "-95,000",
"Dividends paid": "-5,366,000",
"Other financing activities": "27,000",
"Net cash used provided by (used for) financing activities": "-934,000",
"Net change in cash": "-827,000",
"Cash at beginning of period": "23,117,000",
"Cash at end of period": "23,005,000",
"Operating cash flow": "1,109,000",
"Capital expenditure": "-980,000",
"Free cash flow": "129,000"
},
"29/06/2019": {
"Net income": "8,571,000",
"Other working capital": "17,446,000",
"Net cash provided by operating activities": "18,086,000",
"Investments in property, plant and equipment": "-640,000",
"Acquisitions, net": 0,
"Purchases of investments": "72,000",
"Net cash used for investing activities": "983,000",
"Debt repayment": "-76,384,000",
"Common stock issued": "22,000",
"Common stock repurchased": "-93,000",
"Dividends paid": "-6,853,000",
"Other financing activities": "-458,000",
"Net cash used provided by (used for) financing activities": "-25,739,000",
"Net change in cash": "-6,670,000",
"Cash at beginning of period": "23,005,000",
"Cash at end of period": "17,010,000",
"Operating cash flow": "18,086,000",
"Capital expenditure": "-640,000",
"Free cash flow": "17,446,000"
},
"29/06/2020": {
"Net income": "9,634,000",
"Other working capital": "37,268,000",
"Net cash provided by operating activities": "38,860,000",
"Investments in property, plant and equipment": "-1,592,000",
"Acquisitions, net": "-18,000",
"Purchases of investments": "-18,000",
"Net cash used for investing activities": "3,696,000",
"Debt repayment": "-67,532,000",


5 DBS
7 COLLECTIONS
FAVORITE
HOST
localhost:27017
CLUSTER
Standalone
EDITION
MongoDB 5.0.3 Community
Filter your data
Stocks_db
balance_sheets
cash_flow
income_statements
stock_history_average
summary
admin
config
local
Stocks_db.income_statements
Documents
Aggregations
Schema
Explain Plan
Indexes
Validation
FILTER
{ field: 'value' }
ADD DATA
VIEW
{
"_id": {
"$oid": "617a135f822b350f31e3ac4f"
},
"CBA": {
"29/06/2018": {
"Total revenue": "25,709,000",
"Selling general and administrative": "1,159,000",
"Total operating expenses": "10,776,000",
"Interest expense": "16,202,000",
"Income before tax": "13,420,000",
"Income tax expense": "4,026,000",
"Income from continuing operations": "9,394,000",
"Net income": "9,329,000",
"Net income available to common shareholders": "9,329,000",
"Basic EPS": "5.34",
"Diluted EPS": "5.18",
"Basic average shares": "1,746,000",
"Diluted average shares": "1,852,000"
},
"29/06/2019": {
"Total revenue": "24,546,000",
"Selling general and administrative": "943,000",
"Total operating expenses": "11,922,000",
"Interest expense": "16,468,000",
"Income before tax": "11,763,000",
"Income tax expense": "3,391,000",
"Income from continuing operations": "8,372,000",
"Net income": "8,571,000",
"Net income available to common shareholders": "8,571,000",
"Basic EPS": "4.86",
"Diluted EPS": "4.69",
"Basic average shares": "1,765,000",
"Diluted average shares": "1,897,000"
},
"29/06/2020": {
"Total revenue": "24,185,000",
"Selling general and administrative": "828,000",
"Total operating expenses": "11,290,000",
"Interest expense": "11,552,000",
"Income before tax": "10,479,000",
"Income tax expense": "3,020,000",
"Income from continuing operations": "7,459,000",
"Net income": "9,634,000",
"Net income available to common shareholders": "9,634,000",
"Basic EPS": "5.45",
"Diluted EPS": "5.23",
"Basic average shares": "1,768,000",
"Diluted average shares": "1,895,000"
},
"29/06/2021": {
"Total revenue": "24,085,000",
"Selling general and administrative": "940,000",
"Total operating expenses": "11,746,000",
"Interest expense": "5,819,000",
"Income before tax": "12,375,000",


5 DBS
7 COLLECTIONS
FAVORITE
HOST
localhost:27017
CLUSTER
Standalone
EDITION
MongoDB 5.0.3 Community
Filter your data
Stocks_db
balance_sheets
cash_flow
income_statements
stock_history_average
summary
admin
config
local
Stocks_db.stock_history_average
Documents
Aggregations
Schema
Explain Plan
Indexes
Validation
FILTER
{ field: 'value' }
ADD DATA
VIEW
{
"_id": {
"$oid": "617a5af5017aaa7d04ae8bc8"
},
"CBA": {
"1991-10": {
"Open": 6.79,
"High": 6.79,
"Low": 6.79,
"Close": 6.79,
"Adj Close": 1.34,
"Volume": 0
},
"1991-11": {
"Open": 7.51,
"High": 7.51,
"Low": 7.51,
"Close": 7.51,
"Adj Close": 1.48,
"Volume": 0
},
"1991-12": {
"Open": 7.41,
"High": 7.41,
"Low": 7.41,
"Close": 7.41,
"Adj Close": 1.46,
"Volume": 0
},
"1992-01": {
"Open": 7.87,
"High": 7.87,
"Low": 7.87,
"Close": 7.87,
"Adj Close": 1.56,
"Volume": 0
},
"1992-02": {
"Open": 7.33,
"High": 7.33,
"Low": 7.33,
"Close": 7.33,
"Adj Close": 1.45,
"Volume": 0
},
"1992-03": {
"Open": 7.45,
"High": 7.45,
"Low": 7.45,
"Close": 7.45,
"Adj Close": 1.47,
"Volume": 0
},
"1992-04": {
"Open": 7.36,
"High": 7.36,


5 DBS
7 COLLECTIONS
FAVORITE
HOST
localhost:27017
CLUSTER
Standalone
EDITION
MongoDB 5.0.3 Community
Filter your data
Stocks_db
balance_sheets
cash_flow
income_statements
stock_history_average
summary
admin
config
local
Stocks_db.summary
Documents
Aggregations
Schema
Explain Plan
Indexes
Validation
FILTER
{ field: 'value' }
ADD DATA
VIEW
{
"_id": {
"$oid": "617a132e822b350f31e3ac48"
},
"CBA": {
"Value": {
"Previous close": "106.10",
"Open": "105.99",
"Bid": "106.12 x 36400",
"Ask": "106.14 x 7100",
"Day's range": "105.57 - 106.22",
"52-week range": "66.04 - 109.03",
"Volume": "766174",
"Avg. volume": "2841084",
"Market cap": "188.143B",
"Beta (5Y monthly)": "0.64",
"PE ratio (TTM)": "19.66",
"EPS (TTM)": "5.40",
"Earnings date": "10 Aug 2021",
"Forward dividend & yield": "4.00 (3.81%)",
"Ex-dividend date": "17 Aug 2021",
"1y target est": "95.82"
}
}
}
{
"_id": {
"$oid": "617a133a822b350f31e3ac4a"
},
"ANZ": {
"Value": {
"Previous close": "28.39",
"Open": "28.50",
"Bid": "28.54 x 190000",
"Ask": "28.54 x 341400",
"Day's range": "28.45 - 28.98",
"52-week range": "18.52 - 29.64",
"Volume": "4699525",
"Avg. volume": "4569014",
"Market cap": "81.24B",
"Beta (5Y monthly)": "0.90",
"PE ratio (TTM)": "17.33",
"EPS (TTM)": "1.65",
"Earnings date": "27 Oct 2021",
"Forward dividend & yield": "1.05 (3.70%)",
"Ex-dividend date": "10 May 2021",
"1y target est": "29.62"
}
}
}